Human Translation Quality Estimation

Feature-based and Deep Learning-based

Yu Yuan

Supervisor: Dr Serge Sharoff
Dr Bogdan Babych

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

The University of Leeds
Centre for Translation Studies

SLCS August 2018

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. Where other sources of information have been used, they have been duly acknowledged. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in work which has formed part of solely or jointly authored publication. The contribution of the candidate and other co-authors to this work has been explicitly indicated below.

'Section 3.3 of Chapter 3 Features for Human Translation Quality Estimation' contains work which has been partially published in the jointly authored publications:

Yuan, Y., Sharoff, S. (2018) Investigating the Influence of Bilingual MWU on Trainee Translation Quality. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 7-12 May 2018, Miyazaki,Japan.

Yuan, Y., Sharoff, S. and Babych, B. (2016) MoBiL: A hybrid feature set for Automatic Human Translation quality assessment. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pp. 3663-3670, 23-28 May 2016, Portoroz, Slovenia.

'Chapter 4 Cross-lingual Terminology Extraction for Human Translation Quality Estimation' contains work which has been partially published in the jointly authored publications:

Yuan, Y., Gao, Y., Zhang, Y., and Sharoff, S.(2018) Cross-lingual Terminology Extraction for Translation Quality Estimation. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 7-12 May 2018, Miyazaki,Japan.

Yuan, Y., Gao, J., Zhang, Y.(2018) Supervised learning for robust term extraction. In: Proceedings of the 2017 International Conference on Asian Language Processing (IALP 2017), pp.302-305 5-7 December 2017, Singapore.

The author confirms that he is the lead, primary and corresponding author directly responsible for designing the above work, processing data, running experiments, preparing the manuscripts and handling revisions. The co-author Dr Serge Sharoff is his main supervisor, who provided the intellectual input and feedback for data analysis. The co-author Dr Yue Zhang hosts his research in Singapore as a research associate. Other co-authors have kindly offered many suggestions on improving the manuscripts in presentation format and phrasing.

Yu Yuan
August 2018

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisors Dr Serge Sharoff and Dr Bogdan Babych. I owe my deepest gratitude to Serge for his patience, motivation, immense knowledge and continuous support of my PhD and related research. He has introduced me to the world of Computational Linguistics and Machine Learning. His guidance helped me through the research and writing of this thesis, allowing me to have the full academic freedom to follow my ideas and interests. I am also grateful to Bogdan, whose supervision has been equally essential.

I would like to extend my sincere gratitude to a bunch of people who have made the past months of my PhD life smoother and more memorable. I could not be more grateful to Dr Yue Zhang who has hosted my research in Singapore for his generosity. I thank my colleagues (in no particular order) Vivian, Jun, Linda, Long, Jiachen, Valentina, Beatrice, Sophiko and many others who I cannot name here one by one in Michael Sadler B01 & B02 at Leeds for the stimulating discussions, encouragement and all the fun we have had in the past years. Friends in Singapore, Jie, Zhiyang, Fei, Deng, Qi, Shaolei and Yuze, have been very helpful in discussions of coding and debugging. Also, I feel grateful to my colleagues and friends back in China. Their help, whether great or small, is always on my mind.

Last but not the least, I would like to thank my family. My wife has been unwaveringly supporting me throughout my academic career, taking the main responsibility of raising two lovely boys and looking after me. I feel indebted to the boys Miller and Logan for having deprived them of so much daddy-son playing time. The unconditional love and support of my parents and brothers have been my spiritual pillars throughout writing this thesis and my life in general.

This work is under the sponsorship of CSC-Leeds Joint Scholarship (Scholarship Reference:Liujinfa[2013]3009).

Abstract

This thesis studies the technical and linguistic aspects of human translation quality estimation (HTQE) for trainee translations from English to Chinese. To this end, it is cast as a supervised machine learning task through conventional feature-based learning and deep learning to predict fine-grained translation quality scores through regression, using no reference translations.

I investigated how human translations (HTs) can be effectively represented at both the document-level and the sentence-level for quality estimation, exploiting feature-based and deep learning-based methods. Specifically, an extensive framework of translation quality features has been designed at both the sentence- and document-level, and a novel stacked neural model with a cross-lingual attention mechanism, leveraging the strengths of convolutional neural networks and recurrent neural networks, also has been proposed.

From the feature-based perspective, a supervised classification method is proposed to identify terminology for quality evaluation purpose, using language-independent statistics as features. I investigated the correlation of normalised term occurrences with human annotated quality scores. Descriptive and exploratory statistics are carried out on trainee and machine translation datasets through pairwise correlation and principal component analysis to study the contribution of individual and group features and the distribution of translation errors, having shown that HT errors cause mainly content inadequacy and machine translation (MT) errors are more about language misuse. Fine-grained document-level and sentence-level HTQE models are trained using the state-of-the-art XGBoost algorithm with grid search parameter optimisation. Multiple models built with different feature selection strategies are compared to a strong baseline **QuEst** for machine translation quality estimation. On HT and MT data, the optimal models outperform the baseline and other models in predicting the majority of quality scores on the criterion of the agreement with human judgements. From the deep learning-based perspective, a stacked neural model specifically for sentence-level HTQE is presented. The neural architecture has achieved good correlations with human judgements for HTs. For the prediction of MT post-editing efforts, it has achieved comparable performance to a strong baseline for predicting HTER scores of German-English MTs and English-German machine translations (MTs) on the WMT17 test data. The model has also produced good results for predicting keystrokes.

I conclude that this work has created a framework for document-level and sentence-level HTQE and has possibly started a new direction for human translation quality assessment in Translation Studies. The results on HT data show promising performance of the proposed HTQE methods in predicting fine-grained translation quality from multiple aspects, shedding new light on this challenging but essential task in Translation Studies and NLP.

Contents

List of Figures

List of Tables

Abbreviations and Symbols

Roman Symbols

C	a constant integer for normalisation

Greek Symbols

β	a penalty weight, e.g. in BLEU and GlossEx
χ	chi-square statistic
ρ	Spearman Rank Correlation Coefficient
r	Pearson Correlation Coefficient
τ	Kendall's Rank Correlation Coefficient

Other Symbols

$\cdot$	multiplication
$\bigcap$	intersection
$\prod$	dot product of all variables
$\sum$	summation of variables
$\bigcup$	union

Acronyms / Abbreviations

BiLSTM	Bidirectional Long Short Term Memory
BLEU	Bilingual Evaluation Understudy
CBD	City BLock Distance
CDER	Cover Disjoint Error Rate
CNN	Convolutional Neural Networks
CTB	Chinese TreeBank
DL	Deep Learning

DTP	Direct Translation Probability
EDU	elementary discourse unit
EMD	Earth Movers Distance
EM	Expectation-Maximization
GM	Geometric Mean
HT	human translation
HTQE	Human Translation Quality Estimation
ITG	Inversion Transduction Grammar
ITP	Inverse Translation Probability
LDA	Latent Dirichlet Allocation
LSP	Language Service Providers
METEOR	Metric for Evaluation of Translation with Explicit word Ordering
MLE	Maximum Likelihood Estimate
ML	machine learning
MT	Machine Translation
MTQE	Machine Translation Qualty Estimation
MTQE	machine translation quality estimation
MWT	Multi-word Terminology
NIST	National Institute of Science and Technology
NLP	Natural Language Processing
NMT	Neural Machine Translation
PCFG	Probabilistic Context Free Grammar
PMI	pointwise mutual information
PTB	Penn TreeBank
QE	Quality Estimation
QuEst	an open source software is aimed at quality estimation
RNN	Recurrent Neural Networks

RST	Rhetorical Structure Theory
RTM	Referential Translation Machine
SL	source language
SMT	Statistical Machine Translation
SR	semantic role
ST	source text
SVM	Support Vector Machine
SWT	single word terminology
TER	Translation Edit Rate
TF-IDF	term frequency–inverse document frequency
Tree-LSTM	Tree Long Short Term Memory
TL	Target Language
TQE	Translation Quality Evaluation
TS	Translation Studies
TT	target text
WER	Word Error Rate

Chapter 1

Introduction

1.1 Introduction

The notion of quality plays a central role within and outside translation studies. **Translation quality evaluation** (TQE) takes various forms. It either forms part of the formative assessment in education programs at the initial acquisition levels or develops instruments and processes for measuring the quality of translation for certification and research purpose at more advanced levels (Angelelli and Jacobson, 2009a). **Language service providers** (LSPs) also relies on TQE for quality control and assurance. TQE can be performed automatically by computer programmes and manually by human experts. The automated TQE (ATQE) works on both MTs and HTs, including **machine translation quality estimation** (MTQE) and **human translation quality estimation** (HTQE). MTQE predicts the predefined quality labels (i.e. scores or classes) of the unseen, new MTs, as a counterpart to HTQE which predicts the quality of HTs. The researcher can envisage some practical scenarios where HTQE is advantageous.

HTQE helps screening translators and translations with reduced cost and increased efficiency. Every year a large number of applicants submit their sample translations to universities which offer degrees in translation for application purpose. Scoring these translations does require a lot of human resources and time. Therefore, it would be ideal to have low-quality translations filtered so that the designated evaluators only need to evaluate a shortened list of applications. More importantly, in terms of scalability, large-scale translation certification examinations, such as ATA certification Exam[1], ITI professional assessment[2] and CATTI[3], use HTQE will reduce the cost of organising the examination and mitigate the subjectivity of human evaluation. However, these goals can only be achieved if HTQE systems can reliably assess the quality of human translations.

[1] https://www.atanet.org/certification/aboutpractice_test.php
[2] https://www.iti.org.uk/membership/professional-assessment
[3] http://www.catti.net.cn/

HTQE is also supportive to translation teaching and learning. To help trainee translators improve their translation competence, translation instructors grade translation exercises to provide constructive feedback to those taking a course. Translation knowledge and skills of trainees at the end of a translation course often involve grading translations of varying lengths. For trainee translators and language learners, fine-grained HTQE could be a helpful alternative to their course instructors, who are not always available for consultation. In a word, HTQE could facilitate trainees' autonomous learning by providing feedback to their translation exercises so that they can improve through self-reflections and in-depth diagnosis analyses.

Besides, HTQE helps promote translation services and elevate the professional service standards. It is common to carry out the quality assessment in the industry to ensure service quality. Certifying professional competence in translation is one of the means to elevate professional standards, enhance individual performance and identify translators who demonstrate the required professional competence (translation knowledge and skills) to deliver quality translation. Fast turn-around of quality evaluation is also desirable for quality assurance and control. Segments or whole documents that could not pass the predefined threshold will be highlighted and returned to the responsible translators or reassigned to a more competent translator. For translation or localisation service users who do not always possess a working bilingual proficiency, they need to have some professional opinions on their side to determine the worthiness of the service they purchased. Nevertheless, such expert input may not be immediately available unless a reliable HTQE system is accessible to them.

However, we all are aware that HTs are often manually evaluated by qualified and experienced translators and well-trained language tutors. As the prevalent form, manual evaluation is accurate but suffers the deficiencies of long time and high cost. For example, the cost of hiring skilled translators to translate the CHI proceedings (1982-2011) is roughly $\$2.2$ million (Green et al., 2013), and the cost of quality evaluation amounts to at least one-third of this cost. In general cases, it would take 1-2 hours for an experienced translator to review and post-edit a non-domain-specific document of 3 pages. Alongside the issue of time and cost, subjectivity is another problem. Different evaluators may assign different quality scores and classes to the same translation, and the same evaluator will give a different rating to the same translation he or she has just reviewed. In contrast, ATQE partially overcomes the shortcomings of low-efficiency, expensiveness and inconsistency of manual evaluation. It is often cast as a quality estimation task using machine learning (ML) to predict translation quality of labels and/or scores in a given range without the use of reference translations.

However, existing research on ATQE is mainly about MTQE and HTQE is apparently under-researched (Yuan et al., 2016). To this end, I study HTQE, particularly for trainee translations in this work. I investigate how lexical, semantic, syntactic and discourse-level features can be exploited to represent HTs. I explore feature-

based and representation-learning based methods of quality estimation of varying granularities, e.g. at the sentence-level and document-level, and of different quality aspects, e.g. from fine-grained quality subscores to their weighted summation.

1.2 Motivation

Nowadays, artificial intelligence has become an integral part of our life. We are surrounded by smart devices and systems which have brought us the greatest ever convenience and productivity. These applications have maximally emancipated us from the tedious, laborious tasks. Highly intellectual work, such as question answering, translation, essay scoring, can be done by computers.

Using computers to evaluate translation quality is not a totally new task. In the field of **machine translation** (MT), it is indeed quite popular. Reference-based and reference-free MTQE are two common approaches to MT evaluation. The former often refers to reference-based MT evaluation metrics, such as BLEU (Papineni et al., 2002), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005), as these metrics require a corpus of quality human reference translations for comparison to measure the closeness to target translations. The latter, also known as MTQE, demands no reference translations but is able to induce quality scores and labels for new, unseen translations based on a proportion of annotated training data. MTQE acknowledgedly treats the quality evaluation as a prediction task where reference-independent features are extracted from the input source sentences (texts) and their corresponding translations (Specia et al., 2010). MTQE methods have shown their advantages of evaluating MT output quality at run-time in previous work. Findings of these studies have shown that good correlations with human judgements on quality are achievable (Bojar et al., 2016b, 2017).

While quality estimation has been a topic of increasing interest in MT, quality evaluation for HTs is overwhelmingly performed by qualified evaluators who are experienced translators, certified translation reviewers, and translation course instructors. There exist a plethora of translation theories such as skopos theory and post-structural theory for manual quality assessment. In **Translation Studies** (TS), different schools of translation models devote themselves to assumptions about the relationship between the original text and its translation, how the original text is perceived by the author, the translator and the recipients and the cultural and socio-economic consequences which translations have in comparison to other types of multilingual text production. Thus, a typical way of manually evaluating a translation is to make assumptions about the constraining factors for a translation (e.g. purpose, target recipients, budget, target domain) and discuss how the translators have adequately taken into consideration such factors when they are translating. House (2014) systematises and categorises these intellectual approaches as four main groups, namely psycho-social approaches, response-based approaches, text- and

discourse-oriented approaches and philosophical and socio-cultural approaches. Overall, these **intellectual approaches** to translation criticism and assessment are useful for pedagogical and research purpose.

However, for large-scale or fast quality evaluation, these methods are expensive to implement due to the prerequisite of bilingual proficiency and labour-intensive analysis process, not to mention that some essential parameters proposed by these models, such as 'patronage' (See discussions in Munday, 2016a, pp. 203) and 'in-betweenness' (See discussions in Munday, 2016a, pp. 212), are often ill-defined. As a result, carefully designed **scoring rubrics or guidelines** such as American Translators Association Certification Programme rubric for grading[4] and some programme-specific and university-specific scoring metrics for translation quality evaluation are favoured in practice. Notably, in the translation industry, generic industry standards, e.g. ISO 9000 series, DIN 2345, ASTM F2575-06 Standard for Quality Assurance in Translation, have been developed and used in heavily customised forms by LSPs for quality assessment (Drugan, 2013). New industry standards keep emerging. Among them are the influential Multidimensional Quality Metrics (MQM) (Lommel et al., 2014a), TAUS Dynamic Quality Framework (DQF) (Görög, 2014a; Görög, 2014b) and the latest harmonisation of both MQM/DQF[5]. In general, these efforts aim to standardise the translation evaluation and complement the current evaluation methods with efficiency in minds (Melby, 2015).

Even though the above-mentioned standards and quality evaluation frameworks are necessary to fill the gap between theory and practice (Görög, 2014a), measuring and tracking translation quality is still carried out mostly manually, and thus the issues of cost and efficiency are not well solved. With HTQE, the situation could be improved. The task of HTQE shares some common ground with MTQE, but we must acknowledge that the uniqueness of human translation evaluation may hinder a direct application of MTQE models to HTQE:

Human and machine translations are generally different in errors they contain. For instance, Vilar et al. (2006) carried out human error analysis on three **statistical machine translations** (SMTs), and found that missing words, word order and incorrect words are the top three errors commonly seen in different language directions (English-Spanish and Chinese-English), while for HTs, undertranslation (a translation is less specific and incomplete than the original) and awkwardness (a translation is presented with an awkward style due to word order) and syntactic issues may be more representative (See the discussion in Section 5.4). In a similar study, Ahrenberg (2017) found that the most frequent types of edits necessary to enable publication quality to a Google **Neural Machine Translation** (NMT) are 'word edit', 'form edit', and 'order edit' (Vilar et al., 2006) that include replacing a

[4] http://www.atanet.org/certification/aboutexams_rubic.pdf
[5] http://www.qt21.eu/mqm-definition/definition-2015-12-30.html

content word or phrase, changing word morphologically and reordering so as to be accurate in transferring meaning, in comparison to the HT[6].

Human translations and MT translations evaluation generally work with different translation units. MT systems work on the sentence-by-sentence basis due to the constraints of both modelling and computational complexity (Smith et al., 2015). As a consequence, MTQE is more often carried out at or below the sentence level. This can be evidenced by the fact that the established and well-accepted metrics are working on sentence targets. While MTQE is often isolated from the context without the consideration of information beyond it, human translation evaluation mostly works with documents instead. Thus, going beyond the sentence level seems more natural for HTQE. Even in MT, this research question has drawn attention from researchers in recent years. Conventional MT systems are developed and tuned on a parallel corpus of sentence pairs, and evaluation of their outputs is also restricted to the sentence level. In a large-scale survey, Li (2006) report that as the most popular form of testing, trainees are required to translate an entire text or several passages of a document because context is believed to be more critical in translation, and teachers generally believe translating passages and/or documents can better measure trainees' mastery of translation proficiency than translating a series of de-contextualised sentences or phrases. In the LSP sector, it is also uncommon to see clients bid for individual sentences. Instead, passages or longer documents are committed to being translated as a whole. Thus, while MTQE is dominated by sentence-level quality estimation, HTQE often takes place at both document-level and sentence-level, perhaps giving more prominence to the former.

HTs and MTs opt for different evaluation methods. MTQE predicts *fluency*, *adequacy*, and Human-mediated Translation Edit Rate (HTER) (Snover et al., 2006; Specia et al., 2009a) in order to monitor the development of MT systems, or select candidate translations for use or further post-editing in a heuristic manner. Human translation evaluation tend to evaluate translation proficiency, diagnose particular areas of strength and weakness in translation proficiency, and assess the learning achievement, as well as motivating trainees in the translation learning endeavour (Li, 2006) through **holistic scoring** and/or **analytic scoring**. The former method, using experienced readers who are well trained with bilingual knowledge, treats each translation as a single entity and scores them. The holistic scoring can be very reliable (Huot, 1990), as an economical, flexible and applicable instrument for direct assessment of human written texts such as writing and translation (Veal and Hudson, 1983). The latter quantifies multiple aspects of a task and scores them separately (Zhang et al., 2015). These two methods serve different purposes. Holistic evaluation is useful for quick determination of the worthiness of the translation, and analytic evaluation provides detailed feedback on specific errors or suggests concrete remedial action (Lommel, 2018). For translation, aspects such as content,

[6]The author compared one HT and MT to the same source text. The HT is translated and published in a magazine, and it is considered perfect and used for comparison.

style, cohesion, language use, grammar, and target language mechanics are often considered important and separately scored. Many researchers advocate analytic scoring in that they believe it can provide a more objective assessment, especially when multiple raters are involved (Veal and Hudson, 1983). However, research has also revealed that analytic scoring could be disadvantageous since it is infeasible for an evaluator to keep tracking more than one aspect simultaneously, and this repetitive work may impose cognitive burdens (Douglas and Smith, 1997) on evaluators. Therefore, depending on time, resources, purpose and significance of the evaluation, both methods are used interchangeably on a case-by-case basis. For instance, in some large-scale certification translation exams such as ATA certification test, analytic scoring is used, but during a small quiz for translation class, an instructor may use holistic scoring. Both methods are legitimate ways of evaluating translations, and they pose a challenge for HTQE.

Human translators and MT systems employ different translation strategies. In essence, both SMT and NMT are based on modelling large bilingual data or monolingual data (Artetxe et al., 2017), and the translation process is more like a prediction task through either heuristic or end-to-end learning. These working mechanisms bring about predictability and typicality to machine translations. Human translations, in contrast, are enriched with translation strategies that could not be found in MT translations. Ahrenberg (2017) found over 50 strategies in HTs that are beyond the reach of the state-of-the-art Google Translate[7]. These unique strategies for human translators are numerous, e.g. sentence splitting, shifts of functions or category (translating a non-finite clause with a finite clause, a relative clause with a conjoined clause), explicitation (giving specific referent and adding function words) change point of view and paraphrasing. Thus, using current MTQE frameworks does not suffice to capture its diversity brought about by these translation skills.

Therefore, HTQE methods warrant more in-depth investigation. As a standard practice of MT evaluation, the reference-based approach does not suit HTQE. First, human translation tasks are changing and preparing golden references (standard, professional translations) may not be economical and practical. There are other deficiencies as well, such as scores hard to interpret, variational scores with a different number of references. The reference-free approach does not require preparing standard human reference translations and can be customised to predict multiple quality scores instead of the only lexical similarity score (e.g. BLEU, NIST).

This thesis is devoted to HTQE, exploring feature-based and representation learning approaches to this problem. In the meantime, the significance of sentence-level and document-level quality of HTs in evaluation are also considered in a balanced way, given that holistic and analytic assessment are both important means for human evaluation.

[7] https://translate.google.com/

1.3 Aims and Research Questions

1.3.1 Research Questions

Since HTQE is fundamentally a ML task, which consists of three components: representation, evaluation and optimization (Domingos, 2012a). Thus, it is tantamount to choose a viable representation for learning algorithms such that it makes the difference for determining the success of the whole learning task (Bengio et al., 2013). For this reason, much of the effort of machine learning task is spent on the data preprocessing and transformations that produce an effective representation of the target data, as most current learning algorithms are unable to extract discriminative information from the target data (op. cit.). In this sense, feature engineering highlights the necessity of take advantage of human intuition and prior knowledge of a specific domain to compensate this weakness. In the case of HTQE, the concept of HT quality is often implemented through different labels (by ranking or on a continuous scale) and specific representations extracted from the translations. Given the fact that few studies have investigated HTQE in any systematic way (Yuan et al., 2016) and much of the QE research up to now has focused on MTQE in the annual shared tasks organized by WMT (Callison-Burch et al., 2012; Bojar et al., 2013, 2014, 2015, 2016a, 2017), it becomes imminent to investigate how HTs of varying granularities (sentences and documents) can be effectively represented to support learning algorithms for HTQE. Therefore, this thesis focuses on the following questions:

1. **RQ1: How HTs can be effectively represented at the sentence-level for HTQE?**
2. **RQ2: How HTs can be effectively represented at the document-level for HTQE?**
3. **RQ3: How feature-based learning and deep learning are useful to HTQE at different granularities?**
4. **RQ4: To what extent the proposed HTQE method is reciprocal to MTQE?**

1.3.2 Aims and Objectives

The aim of this thesis is to investigate how HTs, specifically trainee translations, can be represented at both the sentence- and document-level for the purpose of HTQE at different granularities.

The objectives of this research are to:

1. investigate and design novel representations for translation knowledge and look for ways of integrating such shallow and deep linguistic information for HTQE. Specifically, the thesis explores in two directions (**RQ1 and RQ2**):

   (a) feature engineering for HTQE
   (b) deep representation learning for HTQE

2. explore the performance of the proposed frameworks at different granularities (**RQ1,RQ2 and RQ3**).
   (a) fine-grained sentence-level HTQE
   (b) fine-grained document-level HTQE

3. construct a corpus with quality annotation for HTQE (**RQ1 and RQ2**).
   (a) translation error annotation
   (b) fine-grained quality annotation at both the document-level and the sentence-level

4. investigate the performance of QE Models for fine-grained quality (ATA, 2011) prediction (**RQ1, RQ2 and RQ3**).
   (a) terminology
   (b) usefulness
   (c) idiomatic writing
   (d) target mechanics
   (e) other combinations (optional)

5. investigate the applicability of HTQE models to MTQE (**RQ4**).
   (a) comparison of MTs and HTs within the same annotation framework
   (b) fined-grained MTQE at the document level
   (c) task-based MTQE at the sentence level

1.4 Main Contributions

The important contributions of this thesis include:

1. construction of corpora for fine-grained HTQE
   (a) English-Chinese corpus of document-level trainee translations with fine-grained quality annotation
   (b) English-Chinese corpus of sentence-level trainee translations with fine-grained quality annotation

2. development of HTQE models through the conventional feature-based learning and deep learning.

   (a) feature-based models for document-level quality estimation
   (b) deep learning-based neural models for sentence-level quality estimation

3. design of a framework of QE features
   (a) lexical tightness as semantic cohesion
   (b) discourse-aware features
   (c) MT back-translation similarity features
   (d) pseudo-reference similarity features
   (e) log ratio of monolingual features
   (f) alignment features
   (g) bilingual terms
   (h) normalised dependency counts
4. statistical analysis of the contribution of features to different translation quality aspects
5. describing and comparing the translation error distribution in MTs and HTs
6. a supervised method of bilingual terminology extraction for HTQE
7. open-sourced tools for HTQE
8. performance evaluation of HTQE models with different learning techniques (change of learning algorithm, parametrisation and feature selection)
9. investigation of the applicability of HTQE models in MTQE tasks

1.5 Structure of the Thesis

Chapter 2 introduces the two mainstream approaches to automatic quality evaluation. I define the problem of quality estimation, introducing reference-based and reference-free methods for automatic evaluation of translation quality.

Chapter 3 proposes a feature set for HTQE and details categories of features and how they are computed from texts.

Chapter 4 focuses on the method for automatic identification of terms from bilingual texts and incorporating the term occurrence information into the QE task. Correlation analysis is carried out with automatically extracted terms on translations from two domains.

Chapter 5 presents a detailed description of all data and resources used in this study. Based on the adapted version MQM-DQF, distribution of translation errors and their interaction with text types are investigated with principal component analysis. The contributions of individual features to each quality aspect are also explored

with a pairwise correlation analysis. Top N contributive features for each quality component by the criterion of linear correlation are listed, followed by discussions.

Chapter 6 presents further empirical results on document-level quality estimation using the proposed feature set. The effectiveness of the feature set on different granularities of texts is examined. Fine-grained quality scores, *Total*, *Adequacy*, *Fluency*, *Usefulness*, *Terminology*, *Idiomatic Writing* and *Target Mechanics*, are able to be predicted with a supervised learning method on the largest human annotated translation data at the document level. The same feature set, with small adaptation, is then applied to a dataset of MT English-Chinese translation pairs.

Chapter 7 proposes a hierarchical neural model for sentence-level quality estimation. The effectiveness of the proposed model is compared with feature-based baselines. This novel method is then applied to large MTQE data for task-oriented QE, e.g. predicting post-editing time in seconds, the number of keystrokes for revision and HTER scores, showing its robust performance across domains and tasks.

Finally, Chapter 8 summarises the achieved results.

Chapter 2

Automatic Quality Estimation: Overview

2.1 Introduction

Human evaluation by well-trained professionals offers insightful judgements on translation quality. However, it admittedly suffers inefficiency, subjectivity and inconsistency. Automatic evaluation tries to emulate human evaluation and is often evaluated against the criterion to what extent it is in agreement with human judgements. In this sense, it is an imperfect substitute for human evaluation.

Research in automatic evaluation has been active, and new metrics and methods are being constantly proposed. Automatic metrics and evaluation methods have advantages over human assessment in that they are generally quick to run and can be used on a large scale with minimal human efforts. The results are reproducible as running the same metrics on the same dataset multiple times will produce identical results (Przybocki et al., 2009b). Also, a well-tuned metric or model can be reused on other evaluations. In contrast, the manual evaluation has been found to be time-consuming, expensive, untunable and nonreproducible (Han and Wong, 2016). Thus, efficiency, reproducibility and reusability, these advantages make automatic evaluation a better alternative for large-scale or fast evaluation, in particular for MTs, which are usually output in large quantity.

Several events promote the automatic evaluation. In particular, the NIST Metrics for Machine Translation Challenge (MetricsMATR) is a biannual evaluation series that focus entirely on MT metrology, advancing innovations in the development of automated metrics (Przybocki et al., 2009b; Callison-Burch et al., 2010). In the metrics and quality estimation (QE) shared task as part of the Workshop on Statistical Machine Translation (WMT) (Callison-Burch et al., 2012; Bojar et al., 2013, 2014, 2015, 2016a, 2017) participating teams each year submit more than dozens of metrics and estimation systems.

Depending on whether these metrics or systems use external resources and what the techniques are, we can roughly classify them into two categories: namely,

reference-based (e.g. Papineni et al., 2002; Banerjee and Lavie, 2005; Snover et al., 2006; Giménez and Màrquez, 2007; Agarwal and Lavie, 2008) and **reference-free** evaluation (e.g. Specia et al., 2009a; Felice and Specia, 2012; Beck et al., 2014, more work can be found in the WMT QE shared task series).

In the following, I will focus on the fully automatic means of translation evaluation and go through the two approaches.

2.2 Reference-based Approach

Automatic metrics are reference-based, working by comparing the lexical similarity of system outputs and human reference translations. Generally, reference-based metrics reward lexical overlapping between candidate translations and a collection of manually prepared reference translations. Over the past two decades, a variety of reference-based evaluation measures have been developed. The main difference between these metrics is the type of measurement they adopt. Among these metrics, some are based on edit distance, such as word error rate (WER) (Nießen et al., 2000), position-independent error rate (PER) (Tillmann et al., 1997) and Translation Edit Rate (TER) (Snover et al., 2006), which measure the good of candidate translation by its (normalized) edit distance (Li and Liu, 2007) to (a) reference translation(s). Some measures compute lexical precision (matching) between candidate translations and reference translation in proportion to the number of common words or n-grams, such as BLEU (Papineni et al., 2002) and National Institute for Standard and Technology (NIST) (Doddington, 2002), while another collection of metrics pay more attention to lexical recall (coverage), such as ROUGE (Lin and Och, 2004), CDER (Leusch et al., 2006), or a balanced consideration of both precision and recall, such as METEOR (Banerjee and Lavie, 2005) and MaxSim (Chan and Ng, 2008). The most widely used MT evaluation metrics in MT literature perhaps are BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), NIST and WER. I briefly introduce the above-mentioned metrics in the following, as in Chapter 3 most of them are used as part of the pseudo-reference and back-translation features.

2.2.1 BLEU

The de facto standard metric BLEU is based on the number of sharing n-grams between the target translation and human reference translation(s) of the same source sentence/text, using different weighting schemes. The basic idea behind the metric is that by counting the number of position-independent matches between the n-grams of the candidate translation and the n-grams of the reference translation(s) a weighted score is then generated for the candidate. The more matches they have,

the closer the candidate resembles the reference translation(s), and thus the higher quality it has.

BLEU metric is essentially precision oriented. It first computes a modified n-gram precision $\mathrm{P_n}$ for any n.

$$\mathrm{P_n} = \frac{\sum_{\mathrm{C} \in \{\mathrm{Candidates}\}} \sum_{\mathrm{n-gram} \in \mathrm{C}} \mathrm{Count_{clip}}\left(\mathrm{n-gram}\right)}{\sum_{\mathrm{C} \in \{\mathrm{Candidates}\}} \sum_{\mathrm{n-gram} \in \mathrm{C}} \mathrm{Count}\left(\mathrm{n-gram}\right)},$$

which counts the maximum occurrences of a word in any single reference translation, clips the total count of each candidate word by its maximum reference count, adds these clipped counts up and have the sum divided by the total number of candidate words. For each candidate translation, the geometric mean, using n-grams up to length N and a positive weight w_n, as the modified precision score is then calculated and multiplied by the result of an exponential brevity penalty score (BP). Let c be the length of the candidate translation and r be the effective reference length. BP is computed as

$$\mathrm{BP} = \begin{cases} 1 & \text{if c > r} \\ e^{(1-\mathrm{r/c})} & \text{if c} \leqslant \text{r.} \end{cases}$$

Then, BLEU score for each candidate translation is obtained as

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{\mathrm{n}=1}^{\mathrm{N}} w_\mathrm{n} \log \mathrm{P_n}\right),$$

By default, N is set to be 4 and the uniform weights w_n equals $1/\mathrm{N}$. Therefore, BLEU metric falls in the range of $\{0, 1\}$. The number of reference translations may have an incremental effect on its final score, i.e. more reference translations per translation could yield a higher BLEU score. BLEU focuses on correlation with human judgements at the corpus level, and scores are not sensitive to differences between versions of MT systems (Banerjee and Lavie, 2005).

2.2.2 METEOR

METEOR is based on an exact word-to-word matching between the candidate translations and one or more reference translations, supporting matching that are identical and/or morphological variants and synonyms of each other. Each possible match is scored by combining unigram precision, unigram recall, and a measure of the degree to which the word order of the candidate translation differs from the reference translation(s). The score for each candidate translation is the best scoring match among all the matches over all references. This maximal-scoring strategy is also used to calculate an aggregate score for the MT system at the corpus level over all candidates. I detail the metric below. **Unigram precision** considers exact

one-to-one matches between words, which takes the form

$$\mathrm{P} = \frac{\mathrm{m}}{w_{\mathrm{t}}},$$

where m is the number of shared words between the candidate translation and the reference translation, and w_{t} is the number of words in this translation. Unigram precision can be interpreted as the proportion of how many words in the translation occur in the reference translation. One variant of the unigram precision is to have all translations and reference stemmed before calculation. **Unigram recall** computes the ratio of how many words in the reference occur in the translation in the formula:

$$\mathrm{R} = \frac{\mathrm{m}}{w_{\mathrm{r}}},$$

where m denotes the number of matching words, and w_{r} represents the number of words in the reference. In the similar vein, a variant of it can be computed by stemming both translations and references. The harmonic mean is then computed:

$$\mathrm{F}_1 = \frac{2\mathrm{PR}}{\mathrm{P} + \mathrm{R}}.$$

To assign more weight to unigram recall over unigram precision, based on the development dataset, FMEAN, as a variant of F_1, is calculated in the form of

$$\mathrm{FMEAN} = \frac{10\mathrm{PR}}{9\mathrm{P} + \mathrm{R}}.$$

METEOR also computes a penalty score for any pair of translation and reference to take into consideration longer matches as

$$\mathrm{Penalty} = 0.5 * \left[\frac{\#\mathrm{chunks}}{\#\mathrm{unigrams\_matched}}\right]$$

where chunks are groups of unigrams in adjacent positions in the translation that are mapped to groups of consecutive unigrams in the reference. Finally, the METEOR score for a translation is given as

$$\mathrm{score} = \mathrm{FMEAN} * (1 - \mathrm{Penalty}).$$

While having demonstrated great promise, the authors claimed that METEOR still has some space to improve. First, the $\mathrm{Penalty}$ score is empirically set on the basis of a test set, and perhaps it would be more optimised to train it on a large dataset and choose a value that best correlates with human judgements. Second, the exact match of unigram could be improved by enabling the metric to match semantic-related words and increasing the coverage of synonyms. In addition, multiple references could have been more effectively used with a synthetic score of all comparisons between references and the translation.

2.2.3 NIST

Similar to BLEU, the NIST metirc is also based on the idea of modified n-gram precision, with some alterations to give more weight to the rarer n-grams (Zhang et al., 2004). NIST also differs from BLEU in term of brevity penalty in that small variational translation length does not impact much the overall score. It can be seen as an upgrade to BLEU. First, the information weightings for n-grams count in the metric are calculated using Equation (2.1):

$$\mathrm{info}(\mathrm{n-grams}) = \log_2 \left(\frac{\mathrm{Count}_{(\mathrm{N-1gram})}}{\mathrm{Count}_{\mathrm{Ngram}}} \right), \tag{2.1}$$

which is part of the overall formula for calculating the NIST score in Equation (2.2):

$$\begin{aligned} \mathrm{Score} = \sum_{\mathrm{n}=1}^{\mathrm{N}} \left\{ \sum_{\text{all n-grams that coocur}} \mathrm{Info}\left(\mathrm{n-grams}\right) / \sum_{\text{all n-grams in the system output}} (\mathrm{l}) \right\} \\ exp\left\{ \beta \log^2 \left[\min \left(\frac{\mathrm{L_{sys}}}{\overline{\mathrm{L}}_{\mathrm{ref}}} \right) \right] \right\} \end{aligned} \tag{2.2}$$

where β is the brevity penalty factor (default 0.5) when the number of words in the system output is $2/3$ of the average number of words in the reference translation, N is set to be 5 and $\overline{\mathrm{L}}_{\mathrm{ref}}$ represents the average number of words in a reference translation over all provided references, while $\mathrm{L_{sys}}$ indicates the number of words in a candidate translation (Doddington, 2002, pp. 141).

NIST metric is found to have more discriminative power than the BLEU metric, but it does not have the same effect as the human judgement in that human evaluators tend to give a higher score to fluent translations, but automatic metrics like NIST do not gain much from longer matched n-grams (Zhang et al., 2004) .

2.2.4 WER

Derived from the Levenshtein distance (Levenshtein, 1966; Kruskal, 1983), word error rate (WER) works at the word level, measuring the difference of across systems and improvement within one system in automatic speech recognition (ASR) and MT. It works by aligning the word sequence with the reference word sequence using dynamic string alignment. WER is defined as the proportion of word errors to the number of words input (Morris et al., 2004). For MT evaluation, it measures the minimal number of insertions, deletions and substitutions necessary to transform corresponding sentences (e.g. target and reference translation) into each other.

$$WER = \frac{\mathrm{S} + \mathrm{D} + \mathrm{I}}{\mathrm{N}_1 = \mathrm{S} + \mathrm{D} + \mathrm{C}},$$

where S is the number of substitutions, D represents the number of deletions and I stands for the number of insertions, with C being the number of corrects. Therefore,

N_1 is the total number of words in the reference. In this sense, WER measures how much cost it takes to get from the target translation to the reference translation through 'deletions' and 'insertions'.

However, edit distance is too simplistic for translations, as legitimate translation variants often differ in word order and WER highly penalises such re-ordering by using wrong words in one place and inserting redundant words at another place (Babych, 2014).

2.2.5 Pros and Cons of the Reference-based Approach

So far I have reviewed some common automatic metrics that are primarily designed to monitor MT system development. The applications of such metrics keep extending to other areas such as summarization (Callison-Burch et al., 2006). Han and Wong (2016) posit that reference-based automatic metrics have been widely used for MT evaluation for being

- fast
- cheap
- tunable
- reproducible

Despite these **advantages** discussed above, there are some **limitations** that hinder its application to HTQE. Though it is possible to obtain parallel data for **source texts** (STs) and **target texts** (TT) in languages under examination, it is not guaranteed that we can acquire the same content as translations that can be used as references, in particular when translation tasks are changing. For example, HTs are working on semantically different texts in different domains at different times. As a matter of fact, the problem of limited coverage of accessible parallel data does exist for MTQE as well. It is considered impractical to prepare reference translations for each translation task for evaluation purpose, as most human translation tasks are one-off. Therefore, reference-dependent metrics do not suit HTQE.

It is also worth noting that there are some fallacies in BLEU and the similar metrics. Koehn (2009) summarizes that the main points of critique are:

- **ignorance of the source-side information.**
  Automatic metrics ignored the source-side information, i.e. the source sentence in particular, all together. Such metrics compare the similarity between the human reference as gold standard reference and the MT output, based largely on n-gram co-occurrence. As a consequence, the ST side information has been neglected.

- **sub-optimality of performance**
  The fact that reference-based metrics mostly use single references often undermines their generalising ability, thus fluctuating performance on different batches of translations. In addition, Length bias might be another issue to some automatic metrics, due to different weighting strategies in their calculation. For instance, **TER** clearly favours short translations as longer sentences contain more insertions, deletions and substitutions.

- **ignorance of the relative relevance of different words.**
  Some words matter more than others. For instance, negation 'not' will totally change the meaning of a sentence, but punctuations are often irrelevant. They are treated equally by these metrics. Babych and Hartley (2004) noticed the lack of a model for the relative importance of matched and mismatched items, and extended the BLEU with frequency weights for lexical items from a human reference corpus, showing significantly higher correlation with human intuitive judgements about adequacy and fluency.

- **addressing not the overall grammatical coherence.**
  N-gram-based evaluation is suspected to bias the metric in favour of phrase-based translations (systems), which are not grammatical at the sentence level. For instance, a candidate translation with a number of matched n-grams in wrong word order will receive the identical BLEU score to the one with the same number of matched n-grams in the correct word order. The example below is an example I intentionally swap the sequences of the first candidate translation at both sides of the empty sign so that two candidate translations have the equal number of 1-4 grams and candidate 2 is made ungrammatical. The two translations will be scored exactly the same score 0.10 per the equation for BLEU above, in which case it is against our intuition. Callison-Burch et al. (2006) also pointed out that n-gram based metrics are biased towards statistical systems. In particular, these metrics will consistently overestimate phrase-based MTs over rule-base MTs.

> **Example: BLEU Scoring**
> ST: " sadly , every year thousands of other people are less fortunate , dying while they wait for suitable organs to be found . "
> candidate 1: 他们 在 等待 合适 <empty> 捐献 器官 中 死去
> gloss: they are waiting suitable <empty> donated organ middle die
> candidate 2: 捐献 器官 中 死去 <empty> 他们 在 等待 合适
> gloss: donated organ middle die <empty> They are waiting suitable
> reference : " 令人可悲 的 是 ， 每年 成千上万 的 人 却 没有 这么 幸运 ， 他们 在 等待 合适 的 捐献 器官 中 死去 。 "
> gloss: " sadly - is , yearly thousands of people but no such luck , they are

> waiting suitable - donated organ middle die . "

- **uninterpretable scores.**
  Metric scores, such as BLEU, depend on many factors, e.g. a number of references, the language pair, the domain. These scores become hard to interpret for intra-system segment evaluation, i.e. comparing translated sentences by a system. Sometimes, post-edited human translations are barely assigned higher scores than MT translations, despite their much higher quality. Thus, the correlation of such metrics with human judgements is an artefact of experiment design (Lommel, 2016), as changing and adding more references could change the BLEU scores dramatically.

Callison-Burch et al. (2006) have shown that BLEU and similar metrics fail to model translation variation and thus higher scores indicate no absolute quality improvement. Contradictions to human judgements are found that highly ranked systems by BLEU are poorly evaluated by human raters, and in the 2005 NIST evaluations on Arabic-English, a post-edited submission by monolingual speakers was only assigned with BLEU scores with small increases but with larger improvements in both fluency and adequacy in the manual evaluation. Thus, it would be inappropriate to use BLEU metrics and the similar for comparing systems that are radically different in architecture (Koehn, 2009). In the case of HTQE, if the styles of translators drastically differ, this bias could be problematic. In addition, automatic metrics are found to be unreliable at the segment-level. Results from these metrics are generally more reliable at the corpus level, but not at the word- and sentence-level. As a consequence, results of evaluating individual translations would be unreliable.

2.3 Reference-free Approach

Reference-free approach to TQE is a newly developed technique to predict quality for unseen translations without reliance on human reference translations.

Methods of this approach rely on features extracted from the source, the translation, or from the translation process (Blatz et al., 2004; Specia et al., 2009a) to build predictive models instead of comparing lexically the candidate translations and the prepared references. Reference-free QE has been gaining popularity in recent years, and a series of QE shared tasks have been organised to predict post-editing efforts and/or quality classes of MTs at **word** (Ueffing and Ney, 2005; Luong et al., 2015a; Servan et al., 2015), **sentence** (Quirk, 2004; Gamon et al., 2005; Specia et al., 2009a), and **document** (Soricut and Echihabi, 2010; Scarton et al., 2016; Graham et al., 2017b) level in real time.

Representing translational data for MTQE has been a hot topic. Starting from different approaches, MT researchers use a variety of internal, i.e. MT-system-

based features such as N-best lists[1], alignment table, and external features, i.e. those generated from external linguistic knowledge sources and tools, such as Part-of-Speech (POS) taggers, syntactic parsers.

In the following, I will focus on the feature engineering part of previous research in MTQE so that we can have an understanding of how MT researchers deal with the representation of translations under various learning constraints. Findings from this review are beneficial to the goal of building a more intuitive HTQE framework. Though this thesis focuses on the related task of HTQE at the sentence-level and document-level, I also include the discussions of features used for word-level MTQE to help readers grasp a fuller understanding of the reference-free MTQE research.

2.3.1 Quality Indicators for Word Level MTQE

Word-level QE aims to label the MT 'generated words as either correct or incorrect'[2] and 'enables the system to signal possible errors to the user or propose only those words as translations that are likely to be correct' (Ueffing et al., 2003). Some potential uses of word-level quality estimation include: highlighting words that need editing in the post-editing stage and indicating which portion (s) of the sentence is (are) not reliable. As reported in Bojar et al. (2014), word-level quality estimation tasks often rely on manually designed features and exploit system-based features (e.g. word graph, word posterior probability), alignment context features (source and target alignment and neighbours), lexical features (POS tags), syntactic features (constituent related) and semantic features, e.g. WordNet (Miller et al., 1990) senses.

System-based confidence measures, which are based on N-best lists or word graphs generated by an SMT system, are obtained and decided whether they have exceeded the predefined thresholds so that a word can be tagged as 'correct' or 'incorrect', or any other variants of labels (Gandrabur and Foster, 2003; Blatz et al., 2004; Ueffing and Ney, 2005, 2007; Camargo de Souza et al., 2014). Some systems make use of the n-grams (previous n-th to the following n-th tokens) and skip-grams (previous and next token) and treat the word quality prediction as a task of sequence labelling (Han et al., 2013).

Syntactic representations, such as constituency and common cover links[3], are found to be the most discriminative features among many others, such as position, length, form and surrounding contexts (Bicici, 2013; Martins et al., 2016). Language model (LM) based scores, such as word occurrence in multiple translation systems and POS tag-based LM scores are combined with those commonly used lexical (source and target POS tags) and syntactic attributes (constituent label, depth in

[1] A list of top n translations along with their scores.

[2] There are also other binary variants, such as Keep/change, OK/BAD and multi-class variants, such as Keep, Delete or Substitute (Bojar et al., 2013).

[3] A new representation that shares the advantages of both bracketing and dependencies but also has additional properties not shared by either. This concept was advanced by Seginer (2007)

the constituent tree) to build classifiers for word-level quality prediction (Luong et al., 2014; Wisniewski et al., 2014; Shang et al., 2015; Tezcan et al., 2015; Beck et al., 2016).

People have shown a boosted interest in Deep Learning (DL) in recent years. Researchers make efforts to solve the QE at word level with deep neural networks. Shah et al. (2015) tried to use word embeddings as an additional feature for word-level QE with a Support Vector Machine (SVM) (Vapnik, 1998) classifier. Kreutzer et al. (2015) proposed a bilingual DNN model for word QE in which bilingual correspondences are learnt 'from scratch' to train a continuous space deep neural network with distributed word representations (Mikolov et al., 2013c) and then fine tuned for the QE classification task. Online tools, such as MT systems are also used to extract bilingual information (e.g. the relations between the source segment and a given target segment) by obtaining the overlapping sub-segments of the source and translating them into the target language (TL). The same process is carried out for all the overlapping sub-segments of the target, which are translated into the SL. The resulting collection of sub-segment translations are then compared to identify sub-segment correspondences between TT and ST (Esplà-Gomis et al., 2015, 2016).

In nature, the word level QE is modelled as a sequence prediction problem. This task predicts the quality labels for segments at different levels of granularity. Binary classification in the form of 'OK' and 'BAD' (or similar variants) and multi-class classification in the form of post-editing decisions (Bojar et al., 2013) and specified error types in Multidimensional Quality Metrics (MQM)[4] are two representative learning tasks, with possible more refined levels (Bojar et al., 2014, 2015, 2016a). As stated at the beginning of this section, while word-level QE might be useful in several ways for computer-aided translation (CAT) and human post-editing, the task is deemed impractical for human translations for the following reasons: First, manually preparing word level training and testing data from human translations is extremely expensive; Second, predicting the difference between translations and their post-edited versions[5] has imposed an unnecessary restriction on human translators and limited their choices of words. Finally, I suspect that word-level QE could not decently address the problems of synonyms, polysemes and different freedom of word order in TL. The estimation itself may unfairly favour translations close to the post-edited reference in form. Last but not the least, word-level correspondence is often viewed as a consequence of translation incompetence[6]. Note that higher level sub-clause QE, for instance, phrase-level QE has been conducted as well (Bojar

[4] http://www.qt21.eu/launchpad/content/multidimensional-quality-metrics

[5] In word-level MT QE, training and testing are often based on segments automatically annotated for errors using the alignments provided by the TER tool (Snover et al., 2006). In other words, the task is predicting automatically annotated errors.

[6] wordwise and/or other forms of literal translation are often discouraged in translation textbooks and translator training.

et al., 2016a). Approaches to phrase-level QE are more or less the same as the word-level QE[7].

In the following section, I focus on features used in the sentence-level MTQE.

2.3.2 Quality Indicators for Sentence-level MTQE

It is widely recognised that the automatic evaluation of MT quality is crucial for inter-system comparisons in the same translation task and intra-system evaluation during the development phase. Most current QE research is carried out at the sentence level.

There are a number of attempts using machine-learned classifiers for the sentence-level MTQE. In the following, I go through some representative work and the features that are used.

Quirk (2004) uses a small corpus (350 sentences) of human annotated MTs to train a classifier that emulates human scoring. He uses features spanning from sentence perplexity score (3-gram LM based) to source and target sentence length, logical form (LF) representations (a predicate argument structure representation), the number and average size of learned mappings, counts and percentages of words translated and target and source ratios of the monolingual features. Some researchers attempt to automatically distinguish MTs and HTs, i.e. human-likeness classification, using perplexity-based features, linguistic features, n-gram precision, length and word error rate concerning human reference translations (Corston-Oliver et al., 2001; Kulesza and Shieber, 2004). Features used in these work can be adapted for the QE purpose.

Gamon et al. (2005) investigate the possibility of detecting dysfluent MT sentences in the absence of reference translations. Sentences are represented as vectors of binary features based on linguistic analysis toolkit. The features they use are based on work in style classification (Gamon, 2004) and fall into several categories: trigrams of POS tags, context-free grammar productions, semantic analysis features, POS and semantic relationship to the parent node, semantic modification relations. Albrecht and Hwa (2007a,b) apply pseudo-reference-based features that are produced by alternative MT systems to regression algorithms in order to measure the quality of MT output sentences. Specia et al. (2009b) exploit resource-independent and system-independent features with inductive confidence machines to dynamically filter out bad translations under certain confidence thresholds. A number of 'black-box' features have been used in their study. These features are mainly from previous work on confidence estimation and have been used in Specia et al. (2009a). In what follows, I summarise their set of 77 features.

- source & target sentence lengths and their ratios

[7] sometimes through certain adaptation, e.g. phrases are treated as a sequence of words. Thus, phrases can be represented as a combination of word-level features.

- source & target sentence 3-gram language model probability & perplexity
- source & target sentence type/token ratio
- source sentence 1 to 3-gram frequency statistics in a given frequency quartile of a monolingual corpus
- alignment scores for source and target and percentage of different types of word alignment, as given by GIZA++ (Och and Ney, 2003)
- percentages and mismatches of many superficial constructions between the source and target sentences (brackets, quotes and other punctuation symbols, numbers, etc.)
- average number of translations per source word in the sentence (as given by probabilistic dictionaries), unweighted or weighted by the (inverse) frequency of the words
- Levenshtein edit distance between the source sentence and sentences in the corpus used to train the SMT system
- source & target percentages of numbers, content-words and non-content words
- POS-tag TL model, based on the target side of the corpus used to train the SMT system

This feature set then is reused in a series of QE experiments exploring QE correlation with human annotators (Specia et al., 2010), investigating more objective ways of annotation for better indicating post-editing effort (Specia, 2011) and predicting translation adequacy (Specia et al., 2011). Specia et al. (2011) started to incorporate into the feature set linguistic information, for instance, POS tagging, chunking, dependency relations and named entities. In a recent work, Felice and Specia (2012) advanced an extended set of 70 linguistics features, complemented by a set of 77 shallow, non-linguistic features, which are extracted from both STs and TTs and summarized below (S for source and T for target).

- sentence 3-gram log-probability and perplexity using a language model (LM) of PoS tags [T]
- number, percentage and ratio of content words (N,V, ADJ) and function words (DET, PRON, PREP, ADV) [S & T]
- width and depth of constituency and dependency trees for the input and translation texts and their differences [S & T]
- percentage of nouns, verbs and pronouns in the sentence and their ratios between [S & T]

- number and difference in deictic elements in [S & T]
- number and difference in specific types of named entities (person, oragnization, location, other) and the total of named entities [S & T]
- number and difference in noun, verb and prepositional phrases [S & T]
- number of unlinked determiners [T]
- number of explicit (pronominal, non-pronominal) and implicit (zero pronoun) subjects [T]
- number of split contractions in Spanish
- number and percentage of subject-verb disagreement cases [T]
- number of unknown words estimated using a spell checker [T]
- number and proportion of unique tokens and numbers in the sentence [S & T]
- sentence length ratios [S & T]
- number of non-alphabetical tokens and their ratios [S & T]
- sentence 3-gram perplexity [S & T]
- type/token ratio variations[8]
- average token frequency from a monolingual corpus [S]
- mismatches in opening and closing brackets and quotation marks [S & T]
- average number of occurrences of all words within the sentence [T]
- alignment score (IBM-4) and percentage of different types of word alignments by GIZA++

Part of these features have been developed into a strong baseline feature framework using only shallow statistics from the source and target texts and further improved (Specia et al., 2013, 2015) and reused in the consecutive WMT quality estimation shared tasks (Callison-Burch et al., 2012; Bojar et al., 2013, 2014, 2015, 2016a).

Apart from the features listed above, participants in these QE shared tasks and interested researchers often put forward new features with variations on top of the baseline features. For different considerations, these newly proposed features vary in different team submissions and research designs. Sometimes they appear in different names because of the preference of researchers, e.g. pseudo reference may be called as MT output difference or MT consensus in different research (Scarton and Specia, 2014b). Typical among them are back-off n-gram based features,

[8] e.g. corrected TTR (Carroll, 1964), Log TTR (Herdan and Wijk, 1960), Guirauid index (Guiraud, 1954) Uber Index (Dugast, 1980) and Jarvis TTR (Jarvis, 2002)

intra-lingual features[9], cross-lingual features[10] in Langlois et al. (2012), sequential features[11] and syntactic dependency features in Pighin et al. (2012), Almaghout and Specia (2013), and Samad Zadeh Kaljahi et al. (2013), POS N-gram based features in Kaljahi et al. (2014b), Luong et al. (2014), and Tezcan et al. (2015, 2016), edit distance between a translation to the training sentence in Buck (2012), parsing statistics from Probabilistic Context Free Grammar (PCFG) parsing and automatic language quality checking in Avramidis (2012a), Avramidis and Popovic (2013a), Avramidis (2014), and Hokamp et al. (2014), topic-based features using Latent Dirichlet Allocation (LDA) in Blei et al. (2003)[12] and Rubino et al. (2012, 2013a), subsequence-level features[13] in González-Rubio et al. (2012), Okapi BM25 similarity (Robertson and Jones, 1976), i.e. term frequency–inverse document frequency (TF-IDF) in Moreau and Vogel (2012), features from the decoding process[14] in Avramidis (2012b), Soricut et al. (2012), Wu and Zhao (2012), and Rubino et al. (2013b), MT output difference or pseudo reference-based features[15] in Okita et al. (2012), Camargo de Souza et al. (2013), Formiga et al. (2013), and Scarton and Specia (2014b), word alignments in Soricut et al. (2012), Camargo de Souza et al. (2013), and Turchi et al. (2013), style classification features in Moreau and Rubino (2013), semantic role labels in Kaljahi et al. (2014a), monolingual and bilingual word representations in Shah et al. (2015) and Abdelsalam et al. (2016). Lucia Specia keeps a complete list of features used by all participating teams (but 1) in 2012 Workshop on Statistical Machine Translation (SMT) (WMT 12) on her personal webpage[17].

In the following subsection, I look at the features employed in MTQE at the document level.

2.3.3 Quality Indicators for Document-level MTQE

In contrast to sentence-level QE, document-level QE is a relatively under-researched area, with only a few previous studies identifiable in the literature (Soricut and Echihabi, 2010; Scarton and Specia, 2014a; Scarton et al., 2015a,b, 2016; Graham et al., 2017b). As Scarton (2016) discussed, document-level QE is challenging in that assessing document is not as straightforward as assessing words and sentences (i.e. assigning scores to large units of text is extremely difficult because small problems at the word and sentence level interfere in human judgement), and there is little

[9] the average mutual information between words in one sentence.

[10] the average of the mutual information between words in source and target sentences.

[11] n-gram model on different variants of sequence in which non-stop words are replaced with the root of the word, the suffix and the POS of the word.

[12] source and target segment probability distribution over topics for a 10-dimension topic model and cosine distance between source and target topic vectors

[13] frequencies and confidence score computed on the $n-$best translations.

[14] e.g. inverted automatic scores, Mini-/maximal link likelihood

[15] MT outputs are considered iteratively as translation reference and compared to each other using the software TERCOM[16]

[17] http://staffwww.dcs.shef.ac.uk/people/L.Specia/resources/feature_sets_all_participants.tar.gz

parallel data with document-level quality annotations available for training. Since 2015, the quality estimation shared task in WMT workshop series has included document level QE as one subtask (Bojar et al., 2015, 2016a). In the following, I identify a few studies addressing the document-level QE.

Soricut and Echihabi (2010) explore quality assessments on documents for which reference translations are not available, using confidence estimation to predict BLEU scores of the translated documents produced by a given MT system. They use external features that include text-based features (e.g. length of source and target in terms of tokenized words), LM-based features (e.g. document-level perplexity score using 5-gram LM), pseudo-reference-based features (e.g. BLEU scores computed using alternative MT systems' outputs as references), example-based features (e.g. top-100 and bottom-100 development set documents as templates) and training-data-based features (e.g. computing the number of out-of-vocabulary (OOV) tokens). While the researchers report promising results for ranking translations of different source documents and consistent performance across a large variety of languages, they also admit that predicting absolute document-level BLEU scores proved inconclusive.

Soricut and Narsale (2012) recreated the document-level quality prediction based on the predicted sentence-level BLEU-like scores, proposing a new approach of combining sentence-level prediction into the document-level prediction. The same set features as in Soricut and Echihabi (2010) are used.

Scarton and Specia (2014a) hypothesise that features that capture discourse phenomena can improve document-level prediction. They considered discourse features and pseudo-reference features (BLEU-like scores). Discourse features consist of lexical cohesion features, such as average word repetition, average lemma repetition, average noun repetition, and Latent Semantic Analysis (LSA) (Landauer et al., 1998) cohesion features, which are in the form of Spearman rank correlation coefficient (Spearman, 1904) of adjacent sentences and the averaged Spearman rank correlation of all sentences in the document. BLEU and TER scores are computed between the target translations and alternative MT systems at the document level and used as features with baseline features (Specia et al., 2015) and the LSA features to build models to predict BLEU and TER scores for systems of interest. Both LSA features and pseudo-reference features showed improvement over the baseline features. Scarton et al. (2015a) explored document-aware and discourse-aware feature for document-level QE in the shared task. Results show discourse features they implemented in the QuEst framework contribute to the improvement over the baseline. The document-aware features are adapted from the baseline features used for the sentence-level QE (same as in Section 2.3.2). A snippet of the features (mainly discourse-aware features) is provided below.

- word/lemma/noun repetition in the source/target document
- ratio of word/lemma/noun repetition between source and target documents

- number of pronouns in the source/target document
- number of discourse connectives in the source/target document
- number of pronouns of each type according to Pitler and Nenkova (2009)'s classification:expansion, temporal, contingency, comparison and non-discourse
- number of Elementary Discourse Units (EDU)[18] (Mann and Thompson, 1988) breaks in the source/target document
- number of Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) nucleus relations in the source/target document
- number of RST satellite relations[19] in the source/target document

At the WMT16 workshop, two teams participated in the subtask of document-level QE. Scarton et al. (2016) submitted two systems using two different approaches. One uses word embeddings as features and is trained with a Gaussian Process (GP) (Rasmussen, 2004) and the other uses a combination of discourse information and is trained with SVM. Other than all features introduced in Scarton et al. (2015a), entity-graph-based features to measure the coherence between source and target are also introduced in this study. Bicici (2016) reproduced their framework of document-level QE as previously in Bicici (2013) with only small adaptation for different tasks.

Other than these studies directly related to the document-level QE, Graham et al. (2017b) explore the validity of gold standard for document-level QE, investigating to what degree tuning the gold standard impacts the validity of the system estimation performance and proposing direct assessment (DA) on Mechanical Turk as an alternative to reliable and cost-effective gold standard construction.

To conclude, in this section, I have reviewed the research in MTQE at different levels of granularity. Various hand-crafted features and system-dependent features are discussed and presented.

Whereas researchers define features according to the attributes of their proposed features, some try to associate the features with a certain aspect of translation quality. Specia et al. (2011) tried to differentiate adequacy and fluency features with source complexity features. Such features use the source-side information to measure the difficulty of translation. This information is concerned with the lexical complexity (e.g. source type/token ratio) and syntactical depth (e.g. source side syntactic parsing). Fluency features (e.g. n-gram LM perplexity) often gauge how natural and fluent the target text is. Adequacy features link the source and target texts. These features aim to assess the formal correspondence and semantic equivalence between source and target. Examples include simple ratios and frequency counts (e.g. the ratio of the number of noun phrases in the source and the target).

[18] the leaves of a discourse tree that correspond to contiguous atomic clause-like text spans
[19] supporting relations

However, these features were designed in MTQE for the purpose of predicting post-editing effort, post-editing time and other automatic metric scores as a substitute for human scores. In contrast, the working subjects and goal for HTQE are different. HTQE aims to estimate the quality of HTs at the sentence- and document-level without reliance on human evaluators. Quality scores (e.g. on a predefined scale) or labels (e.g. 'good' or 'bad') can be assigned to new translations automatically by the trained system. Thus, the main goal of HTQE is to predict human scores for the unseen translations, instead of metric scores (e.g. HTER and METEOR scores) computed between the references and the translations. These differences require us to rethink what features are suitable to be included for HTQE.

In addition, some fallacies in the current framework of MTQE could be identified. I argue that the practice that MT outputs are mainly evaluated as segments (e.g. at the sentence or sub-sentence level) does not conform to the fact that human translations are often evaluated as a whole piece of work at the document-level through global scoring[20]. Even though QE at the segment level is beneficial for identifying potential problematic words, phrases and sentences in translations, this type of tasks are likely to incur disrespect for discourse-level phenomena in the translation and unfairly favour translation segments that are locally optimal but globally out-of-place. Out of these hundreds of features are only a few crafted for measuring cohesion and coherence of the translations (See my discussion in the last subsection), and also document-level QE is largely neglected in MTQE. In the meantime, it is observed that terminology equivalence, an important factor affecting translation quality (as evidenced by Multidimensional Quality Metrics (MQM)[21] that terminology is one of the main issues), has not been properly tackled. Domain-specific texts contain a number of terms that impose extra cognitive processing load on translators, and the lexical-conceptual representation of those terms may be challenging for them. Thus, whether or not translators can successfully render the terminology in TL is a good criterion for evaluating the quality of target texts (TTs).

2.3.4 Pros and Cons of the Reference-free Approach

In contrast, reference-free QE has demonstrated superiority over automatic metrics. The advantages of the methods of this approach include:

- **minimal human intervention required.**
  It only takes necessary manual labour to annotate enough training data, and it works on unseen new translations without additional manual work, unlike reference-based metrics that can only work for translations of the same content as the reference(s). It requires no preparation of references, which sometimes are costly to obtain, for new evaluation tasks. For instance, for a job of

[20] In global scoring, the examiner reads the entire essay or translation and makes a holistic judgement about the quality (Isenhour and Kramlich, 2008).

[21] http://www.qt21.eu/mqm-definition/definition-2015-12-30.html

approximately 4000 words English-Chinese translation, it will cost a client nearly $\$400$.

- **task-oriented quality evaluation.**
  Different aspects of translation quality in various forms, e.g. from editing cost to adequacy rank, can be its target to estimate. In this study, I estimate quality scores for human translations according to a certain well-accepted instrument such as ATA rubric. The fine-grained quality components of this scheme cannot be obtained from reference-dependent metrics but through reference-free QE techniques.

- **customisable granularities of evaluation.**
  Unlike automatic metrics who yield uniform scores for system-level evaluation, reference-free QE can produce fine-grained scores up to a finite number of levels for words, sentences, and documents.

However, MTQE and HTQE serve different purposes. MTQE at the sentence- or word-level fits best a scenario in which automated translation is only part of a more extensive pipeline. Such pipelines usually involve human post-editing, beneficial to translation productivity (Lagarda et al., 2009). In contrast, HTQE is not designed only for selecting defective translations for post-editing (See my discussion in Chapter 1). QE at the word- and sentence-level, however, suffer from the inherent deficiency that contextual information is often ignored, and deviate from the fact that human translations mostly appear as documents.

In the meantime, I have observed some limitations of current reference-free research:

- **overwhelmingly centred on MTQE.** As I described in the last section, most work focuses on MTQE, and HTQE is clearly under-researched. More efforts should be devoted to it. QE should be tailored to the specificity of HTs, given that they differ in terms of translation errors, working units, evaluation methods and criterion and translation strategies (See my discussion in Section 1.2). Representations for HTs at different granularity levels that facilitate the training and tuning of an effective QE model is of critical significance.

- **no publicly available HTQE datasets.** In contrast to a large amount of MTQE datasets that are accessible to the research community, datasets for HTQE are scarce. To my knowledge, there are no open datasets that have been manually annotated with schemes of translation quality for HTQE research. Therefore, collecting and annotating HTs with a proper quality scheme and making them accessible to the research community are beneficial for further research in HTQE.

- **lack of fine-grained QE.** While evaluating MT from the perspectives of adequacy, fluency and post-editing efforts is typical, I argue that finer-grained

HTQE is more suitable for evaluating human translations, as coarse-grained HTQE cannot provide much insightful feedback regarding linguistic phenomena and translation errors. The granularity of levels is another issue. Current reference-free MTQE mainly works at the sentence-level, leaving document-level QE less-studied. As I previously discussed in Section 1.2, both document-level and sentence-level QE should be considered in a balanced way.

Thus, reference-free approaches in MTQE must be adapted for HTQE by taking into account forms and uniqueness of human translations. Considering the de-facto status of sentence- and document-level MTQE, this thesis aims to carry out a more in-depth study in this direction for HTQE. I aim to address the above-mentioned issues in this study.

2.4 Summary

From a textual-linguistic point of view, a reliable evaluation method should take into consideration the purpose of translation, the end user and text types, among many other factors (House, 2014). Qualified human evaluators are desirable for such a task, but human evaluation suffers its drawbacks of being expensive, time-consuming, and non-reusable. Automatising the evaluation of human translations can mitigate the negative influence of manual evaluation and at least complement human evaluation in certain scenarios, such as low-risk taking examination scoring and autonomous learning feedback.

Reference-based metrics can help overcome the limitations of human evaluation but suffer imperfections making them less desirable for the automatic evaluation of human translations. In comparison, reference-free QE is more task-oriented and capable of estimating fine-grained translation quality at different levels. For instance, a QE system can assess human translation from different aspects of quality.

To sum up, this section introduces two approaches to automatic evaluation of (machine or human) translations. It is argued that reference-free QE is more suitable for automatising human translation evaluation. In the next section, I will focus on the critical part of featuring engineering for HTQE.

Chapter 3

Features for Human Translation Quality Estimation

3.1 Introduction

As discussed in Chapter 1, automatising the process of QE for human translations could be beneficial in many ways. For instance, it can mitigate the negative influence of human subjectivity that is brought about by different quality expectations of the assessors, in addition to the greater productivity it ensures and colossal cost it reduces for evaluating a large number of translations in large-scale exams. However, as Domingos (2012b, pp. 78) noted, much of 'the knowledge that is needed to develop ML applications successfully is often not readily available'. Thus, choosing an intuitive framework of representations for human translations is paramount to successfully building a performant system.

Starting from different approaches, researchers used a variety of internal (i.e. MT-system-based features such as N-best lists[1], alignment table) and external features (i.e. features generated from external linguistic knowledge sources and tools, e.g. Part-of-Speech (POS) taggers, syntactic parsers) for MTQE. We admit that HTs and MTs are similar in that they both are language transfer products, but both types of texts clearly have their own characteristics as human translators and MT systems do not share equally the cognitive capacity and the same working mechanisms (see my discussion in Section 1.2). In the meantime, some fallacies can also be observed in the current MTQE frameworks:

- **MT system-based features are not generalisable to HTQE.** It is natural to utilise internal features extracted from the MT systems to be evaluated. However, some of these features such as N-best list are not easily accessible to HTQE systems. Therefore, they are not generalisable or reproducible in HTQE even though they have proved effective in MTQE tasks.

[1] A list of top n translations along with their scores.

- **Over-reliance on few types of features.** It is noticed that some MTQE research use heavily certain types of linguistic features in building QE models. For instance, Albrecht and Hwa (2007b) rely on pseudo-references to develop sentence-level MT evaluation metrics, and Liu and Gildea (2005) explored the use of kernel-based subtree and headword chain metrics to compute the similarity between target translations and references. Though the effectiveness of these features in QE have been validated, they may have explicitly disregard some useful information.

- **Computational complexity could be further reduced.** Even though computing power for modern computers is no longer a problem, complex ways of computing features for the QE purpose cause problems, e.g. debug, deployment, dependencies, to the QE systems. Take the hand-crafted features in a handful MTQE research for example. Kaljahi et al. (2014b) have designed a hand-crafted set of constituency and dependency related features. Some of these features require additional resources and computation. For instance, the average number of POS n-grams in each n-gram frequency quartile demands treebanks and computing the POS-ngram distribution in different frequency quartiles. To compute dependency relation n-gram scores against language models trained on the respective treebanks for each language, we need an additional language modelling process of dependency relations.

Therefore, in designing the feature set for HTQE, the researcher includes those classic shallow features in previous MTQE research, such as QuEst (Specia et al., 2013) and QuEst++ (Specia et al., 2015). In the meantime, the researcher tries to incorporate multiple categories of features into the framework so as to consider information capturing various aspects of translation quality. Also, simplistic forms of features from STs and TTs (i.e. normalised frequency counts and the ratios between the source and target side feature values) are preferred, in addition to the distances of feature vectors of the same category, e.g. a distance between the two vectors of all POS features in STs and TTs.

Motivated by the reasons above, I intend to develop feature representations for better capturing human translation quality on the basis of the current MTQE framework. I am particularly interested in HTQE for the quality of trainee translations. The main contribution of this chapter lies in the feature set I design, which integrates the massively expanded MTQE features. Novel features such as cohesion and coherence features for the task of HTQE, string- and vector-based pseudo-reference and back-translation features are introduced. The proposed feature set has attempted to extend the features from lexical, syntactic level to discourse level to better adapt to HTs at the document-level. In this chapter, I present the feature set and the intuition behind the design.

3.2 Human Quality Estimation Features

This section will mainly deal with the representation of human translational data for QE purpose. As stated above, the features consist of two components: a portion of features from MTQE research and a newly proposed set.

Translation is a very complex human behaviour, which involves a multitude of factors, such as text types, language pairs, translation tools, deadlines, speed, rates and specifications. Consequently, translation quality is subject to and substantially influenced by inner linguistic-textual factors (e.g. language norms, text types) and extra-linguistic factors (e.g. translation specifications, translator competence) (House, 2014; Munday, 2016b). For simplification, I treat translation as a purely linguistic-textual operation and look into factors at this level only. Translations are viewed as a by-product of monolingual and bilingual communication, constrained by the TL norms. Though the translation process is overwhelmingly black-box in nature, the final product of mental activity is somewhat transparent. In other words, if we are examining the quality of translations from a linguistic perspective, characteristic features for each translation at the lexical, syntactic and discourse level should be the main focus. In the following, I come up with a set of specific features that belong to these categories and discuss their relationship with translation quality.

For presentation purpose, I group them into three main categories of features: monolingual, bilingual and language modelling. For instance, language modelling is monolingual, but features under this category are mainly probabilistic, different from most of the frequency-based features in the monolingual group. For a detailed overview of various features used in MTQE, I refer the readers to the discussion in Section 2.3 and the WMT12 shared task on QE (Callison-Burch et al., 2012). My framework for HTQE provides a wide range of features and methods extracting them from STs and TTs and external resources and tools (Section 5). These features go from simple, language-independent ones to advanced, linguistically motivated ones. Among them, some typical MTQE features from the standard MTQE framework QuEst++ are included in my framework:

- target sentence length in words
- log probability LM for the source
- log probability LM for the target
- perplexity LM for the source
- perplexity LM for the target
- number of sentences in the source
- number of sentences in the target
- number of types in the source

- number of types in the target
- type-token ratio in the source
- type-token ratio in the target
- averaged sentence length in the source
- averaged sentence length in the target
- number of out-of-vocabulary words in the source
- number of out-of-vocabulary words in the target
- number of punctuation marks in the source
- number of punctuation marks in the target
- number of prepositional phrases in the source
- number of prepositional phrases in the target
- (partially) pseudo-references MT metrics

Note that here source and target refer to STs and TTs at both sentence- and document-level. In each group of my framework, all new features I redesign or propose are marked with a triangle.

3.2.1 Monolingual Features

Translation is concerned with two or more language pairs. Monolingual features refer to features that monolingually occur in either TTs or STs. These shallow surface features (e.g. the number of tokens, sentence length) and linguistic features (e.g. POS tags) often present discrepantly in both languages, and their varying distribution has proven contributive for predicting translation quality (Specia et al., 2009b; Callison-Burch et al., 2012). What follows is an account of main monolingual features I have advanced for HTQE.

3.2.1.1 POS Tags

POS tagging is the process of assigning one of the grammatical categories to the given word (Manning and Schütze, 1999; Voutilainen, 2003). Examples of common linguistic categories include nouns, verbs, adjectives, adverbs, prepositions, conjunctions and their subcategories. The distribution of POS is generally seen as a factor strongly related to the syntactic quality.

As linguistically motivated features, POS related features are exploited in Giménez and Màrquez (2007) and Specia et al. (2011) and used as baseline features in the

WMT Quality Estimation shared-task 2012 (Callison-Burch et al., 2012). For instance, number, percentage and ratio of content words and function words are extracted as linguistic features in Felice and Specia (2012). In a similar vein, POS tags were counted as shallow grammatical matches on both the source and the target (Avramidis, 2012b; Beck et al., 2013; Luong et al., 2014). Unlike other shallow features, such as sentence length or n-gram statistics, which are limited in their scope and account for the very superficial aspect of a translation, linguistic features convey meaning, grammar and content (Felice and Specia, 2012). Their dynamic relationship might be contributing to the meaning transfer from STs to TTs. Based on this assumption, I use the POS tagging component from Stanford CoreNLP (Manning et al., 2014) to process STs and TTs and then match them per the Universal POS-Tagset (Petrov et al., 2012–2013) in order to achieve better comparability for a distant language pair, such as the English Penn TreeBank (PTB)[2] and the Chinese TreeBank (CTB)[3]. The universal POS-Tagset has been demonstrated effective in MTQE (Han et al., 2014) and able to deal with incompatibility of two distantly related languages. Chinese has far fewer linguistic categories (34) than Indo-European languages, such as English (45) (Petrov et al., 2012–2013). I count the occurrences of POS tags in STs and TTs according to the converted universal tags.

For a detailed matching between Universal POS-Tagset and the common POS tags in two languages, please refer to Table 3.1

Universal Tag	English	Chinese
	English (PTB)	Chinese (CTB)
.(Punctuation)	! # $ " " ,) -LRB- -RRB- . : ? HYPH	PU
CONJ(conjunctions)	CC	CC CS
VERB(verbs)	VB VBD VBD VBN VBG VBG VBN VBP VBZ VP MD	VA VC VE VV
NOUN(nouns)	NN NNP NNPS NNS NN|NNS NN|SYM NN|VBG	NN NR NT
NUM(cardinal numbers)	CD	OD M CD
PRON(pronouns)	PRP PRP$ PRP|VBP WP WP$ EX	PN
ADJ(adjectives)	JJ JJR JJS JJ|RB JJ|VBG	JJ
ADV(adverbs)	WRB RB|VBG RB|RP RBS RBR RB	AD
ADP(adpositions)	IN IN|RP	P
DET(determiners)	DT EX PDT WDT	DT
PRT(particles or other function words)	RP TO POS	SP MSP LC ETC DEC DEG DEV DT AS
X(foreign words)	UH SYM LS FW	X SB ON LB IJ FW BA

Table 3.1 Universal POS TagSet Mapping from English & Chinese

The distribution of each feature is normalised to take into account the length of text (either sentence or document) for the sake of inter-sentential or inter-document comparison. The normalised feature count is calculated as

$$\mathrm{Freq_f} = \frac{\mathrm{Count_f} \times \mathrm{C}_{[10,100]}}{\mathrm{Len}_{sent/doc}}. \tag{3.1}$$

In Eq. 3.1, $\mathrm{Freq_f}$ stands for the normalised counts for the specific feature in sentences or documents, $\mathrm{Count_f}$ represents the original count, and for presentation

[2]https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
[3]http://repository.upenn.edu/cgi/viewcontent.cgi?article=1039&context=ircs_reports

purposes, $C_{[10,100]}$ denotes a constant number with a binary value of 10 (for a sentence) and 100 (for a document)[4]. $Len_{sent/doc}$ is the length of the text in terms of the number of tokens (words) for either a sentence or a document. This normalisation procedure is applicable to other count-based features hereinafter.

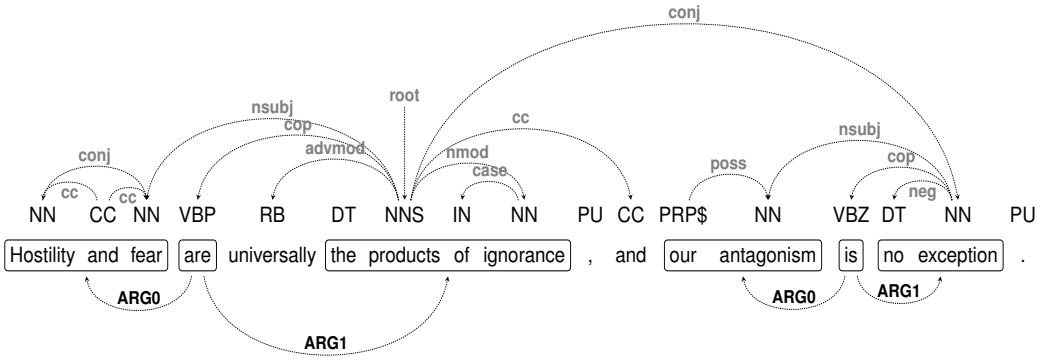
3.2.1.2 Dependency Relations

Typed **dependencies** and phrase structures are different ways of sentence structure representation. While a phrase structure parse represents nesting of multi-word constituents, a dependency parse represents dependencies between words (De Marneffe et al., 2006). Different types of dependency representations acknowledge the semantic, syntactic aspects of texts and are used to parse natural languages.

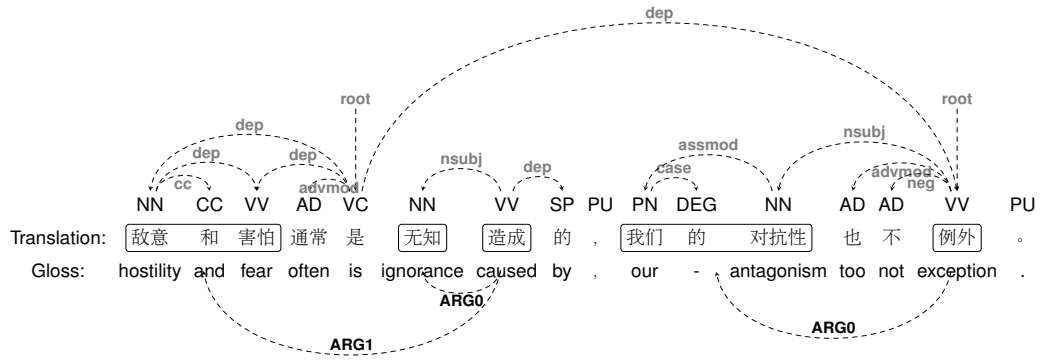
As opposed to the constituency relation, dependency is the notion that linguistic units, such as words, are linked to each other by certain relations and the central verb is at the core of the sentence in which all other elements are under its dominance. Thus, verb dependants are directly or indirectly connected through these dependencies. Each one of them, centring around the main verb (s), constitutes a network of syntactic relations with associated lexical items. This network of typed dependencies (See Figure 3.1a for illustration) challenges the human translators' intellectuality and translation competence. For translators, in the translation process, it requires them to understand and interpret these connected relations and recode them in TL. Therefore, dependency relations in both STs and TTs point a direction for estimating the quality of translations in question.

The dependency tree contains both the lexical and syntactic information, which inspires us to use it for QE. In addition, during the transfer from one language into another, dependencies may, more often than not, demonstrate stability that core verbs and the dependencies under its dominance would reproduce themselves in translations. Intuitively, a good translation normally would have a proportionally equivalent number of dependencies in most cases. For the sentence in Figure 3.1a, I extract two students' translation and plot their dependency relations as well (Figure 3.1b and Figure 3.1c). Parsers are less successful on badly translated sentences, thus leading to mismatched counts of dependencies. As is shown in these figures (note: dependencies are gray labels in Figure 3.1b and red labels in Figure 3.1c), In the translations of the source text sentence, dependencies, such as case (case), nominal subject (nsubj), coordination (CC), negation modifier (neg) and adverb modifier (advmod), are reproduced in both students' translation at exactly the same position. However, the first student's translation (in Figure 3.1b) has yielded several dependencies marked uncertain (dep) due to the literal translation of 'fear' into 害怕 ('afraid'). This mistranslation causes confusion to the parser, thus mismatches in types and numbers of dependencies. In contrast, student 2's translation (in Figure

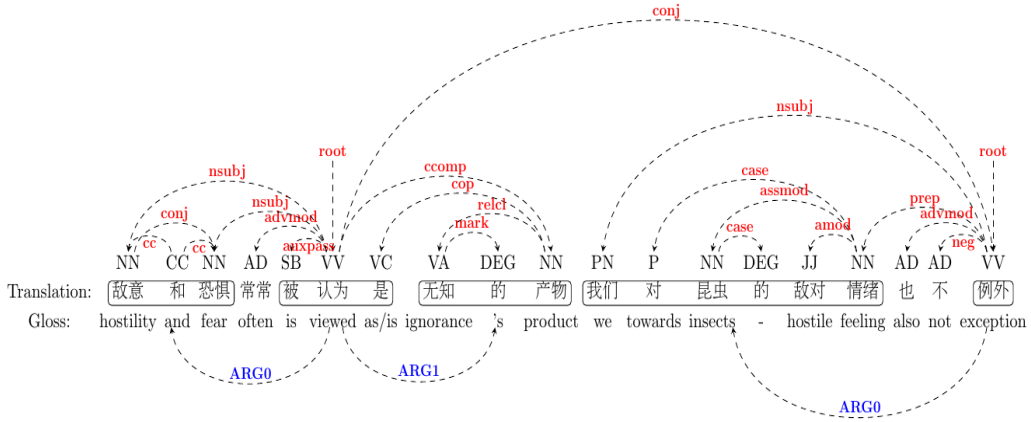
[4] I set this binary value because on average sentences contain fewer than 15 words each, and a translated document has about 250 words each.



(a) Dependencies and Semantic Roles in ST Sentence



(b) Dependencies and Semantic Roles in Translation (Student 1)



(c) Dependencies and Semantic Roles in Translation (Student 2)

Figure 3.1 Dependency and Semantic Parsing Information

3.1c) is less problematic to the parser with more matched dependency relations. In addition, it translates explicitly the part 'our antagonism', which is omitted by the first student.

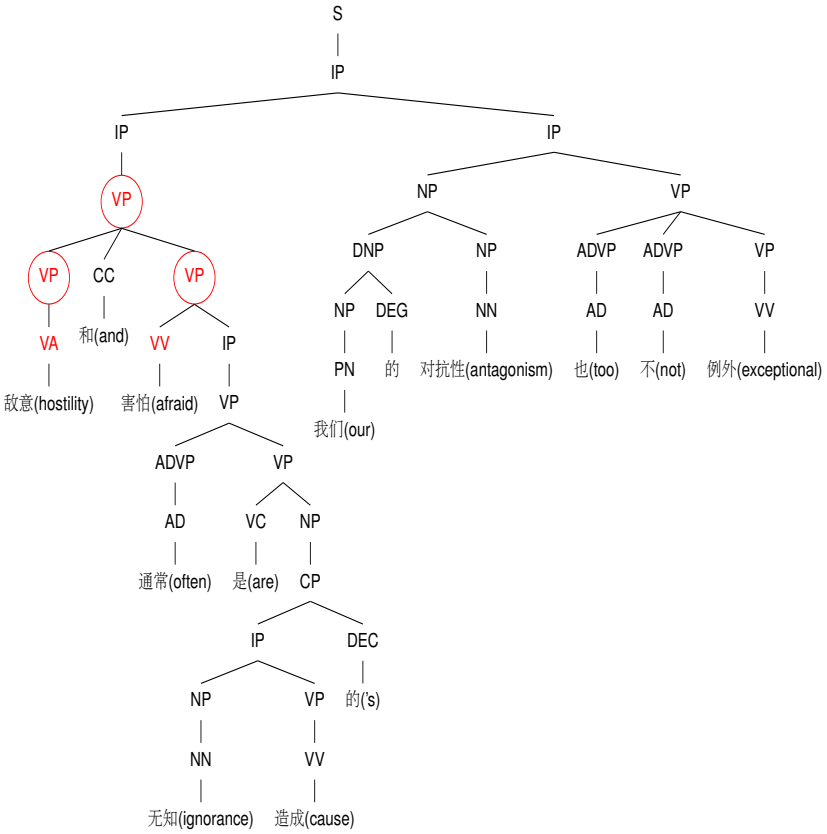
Dependencies have found their way into translation quality prediction in MTQE. For the purpose of evaluating MT outputs, dependencies in the hypotheses are seen as the projection of and compared against the dependencies of the reference translations, and on its basis, their precision and recall of alignment (or other variants) is calculated (Pighin et al., 2012). Instead of computing the accuracy of

the dependency alignments for human translations and references, Kaljahi et al. (2014b) exploit dependency tree kernels in association with hand-crafted syntactic features to predict automatic metric scores. There is no evidence that source-side dependencies will remain intact in the target, especially when two drastically different languages are involved. The researcher takes a more simplistic approach to count the frequencies of each typed dependency in a source and its corresponding target translation and normalise the counts by their lengths respectively. To extract the dependency information, the researcher uses the dependency parser of Stanford CoreNLP that works for both English STs and Chinese TTs. Dependency features are normalised per Equation (3.1). Different from the above-mentioned work in MTQE trying to compare the reference and target translation syntactically (Pighin et al., 2012) or extract syntactic knowledge in a rather complex manner (Kaljahi et al., 2014b), the researcher does not rely solely on syntactic information and instead seeks to integrate it with other potentially useful features. To this end, he associates the dependency relations in STs and TTs with a universal parsing framework (Petrov et al., 2012–2013) for cross-lingual extraction purpose.

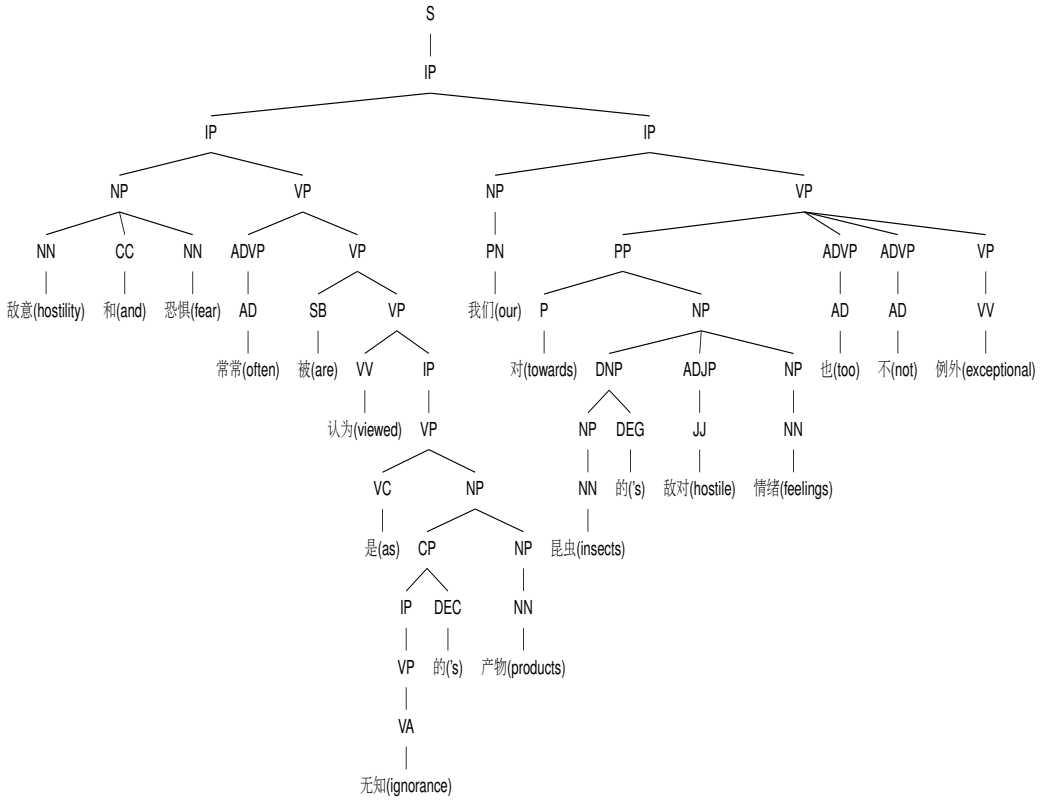
3.2.1.3 Constituency

Apart from dependencies, another very closely related linguistic phenomenon is the phrasal structure in STs and TTs. While dependencies deal with relationships between words, constituencies provide more detailed information about sub-phrases within a sentence. Syntactic features are useful to capture the syntactic complexity of the source sentence, the grammaticality of the target translation and the syntactic symmetry between the source sentence and its translation (Kaljahi et al., 2014b).

Koehn et al. (2003) found that small phrases up to three words could help achieve high accuracy and outperform word-based models in machine translation. Combining fundamental ideas from both syntax-based translation and phrase-based translation, Chiang (2007) used hierarchical phrases in his machine translation model and achieved rather good speed and accuracy. Previous research has also borne out their usefulness in MTQE. Corston-Oliver et al. (2001) classify sentences into MTs and HTs with parse tree features. Quirk (2004) integrates the binary feature of a spanning parse tree with other features for a 4-point scale sentence-level QE. Liu and Gildea (2005), Albrecht and Hwa (2007b), and Giménez and Màrquez (2007) compute the syntactic similarity between MT outputs and reference translations in QE tasks. **Constituency** information in the form of tree kernels (Collins and Duffy, 2001; Moschitti, 2006) collected from the parse trees is now used as features in some QE work (Hardmeier et al., 2012; Rubino et al., 2012), suggesting that correspondence between longer units of STs and TTs might be more associated with quality translation. For human translation, this is related to the 'unit of translation' (Vinay and Darbelnet, 1958) issue.



(a) Constituency Parsing of Student 1's Translation for the Sentence in Figure 3.1a



(b) Constituency Parsing of Student 2's Translation for the Sentence in Figure 3.1a

Figure 3.2 Constituency Parsing Information

Ballard (2010) claims that it is only in the act of translating that units become visible. Therefore, an experienced translator (or a trainee translator) can, judging from his or her past experience or acquired competence, surmise what will constitute a legitimate unit of translation to work in the source text. On this premise, he thinks a unit of translation is generated by the implementation of a translation strategy (literal or non-literal). This view reminds us that translators may approach the ST and the TT in a different manner, as manifested in the length of their working unit particularly. I assume this difference in decision making during the act of translation may lead to variation in units of translation in the form of lexical sequences and syntactic constituencies, which as a result indicate a quality difference.

In Figure 3.2a, three circled nodes are incorrectly parsed because the translator wrongly rendered the NP 'fear' into 害怕 ('afraid'). Student 1 's choice caused confusion to the parser, leading to three parses wrongly annotated as 'VP', including the wrong POS of their children. In contrast, student 2 's translation in Figure 3.2b is almost perfectly parsed, with right constituency for each node. When translating, if Student 1 worked at the NP phrase level, 敌意和恐惧 (hostility and fear) would not have been rendered as two words of different classes. Therefore, translators' choices may contribute to variation in parses that make a difference to the quality of the produced translations. Regarding this factor, main phrasal types are included as one type of monolingual feature in ST and TT.

I extract, process and normalise the counts of constituencies in the same way as for dependencies, using the constituency parser in Stanford CoreNLP.

3.2.1.4 Semantic Role Labels

A **semantic role** (SR) is the underlying relationship that a participant has with the main verb in a clause. SRs are most often embodied by the grammatical relations of the subject, object and indirect object in natural languages. These semantic relations are associated with the agent, force, instrument, experiencer, recipient, and patient in a sentence. Other SRs are more likely to be embodied in oblique (ad-positional) phrases or adverbials (Payne, 1997).

Examples of SRs are numbered arguments[5], adjuncts[6], references[7] and verbs (predicate verbs) (Carreras and Màrquez, 2005). However, due to the systematic variation between two languages, it is often challenging to define a universal set of thematic roles (Jurafsky and Martin, 2009). For instance, the SR labelling in English generally uses the format of annotation in the English Proposition Bank (EPB) (Palmer et al., 2005), and SR labellers for other languages rely on adapted, compatible data formats or proposition banks of their own. For instance, the state-

[5]Verb-specific roles, in general, Arg0 represents the agent, Arg1 the patient, Arg2 often the benefactive, instrument, attribute or end state, Arg3 the start point, benefactive, instrument, or attribute, Arg4 the end point and Arg5 the direction (Bonial et al., 2012).

[6]General arguments that any verb may take optionally.

[7]Arguments realized in other parts of the sentence.

Label Type	English (Roth and Lapata, 2016)		Chinese (Che et al., 2010)	
	PB Label	Description	PB Label	Description
Predicate	rel	verb	root	verb
Numbered Arguments	ARG0 ARG1 ARG2 ARG3 ARG4	semantically licensed by the predicate	ARG0 ARG1 ARG2 ARG3 ARG4 ARG5	verb specific roles
Predicate/phrasal modifiers	ARGM-ADJ	Adjectivals m(modifies nouns)	ADV	adverbial, default tag
	ARGM-ADV	Adverbials (modifies verbs)	BNE	Beneficiary
	ARGM-CAU	Causatives	CND	condition
	ARGM-COM	Comitatives	DIR	direction
	ARGM-DIR	Directionals	DGR	degree
	ARGM-DIS	Discourse markers	EXT	extent
	ARGM-DSP	Direct speech	FRQ	frequency
	ARGM-EXT	Extents	LOC	locative
	ARGM-GOL	Goals	MNR	manner
	ARGM-LOC	Locatives	PRP	purpose or reason
	ARGM-MNR	Manners	TMP	temporal
	ARGM-MOD	Modals	TPC	topic
	ARGM-NEG	Negations	CRD	coordinated arguments
	ARGM-PRD	Secondary Predications	PRD	predicate
	ARGM-PRP	Purpose	PSR	possessor
	ARGM-PRR	Nominal predicates in light verbs	PSE	possessee
	ARGM-REC	Reciprocals		
	ARGM-TMP	Temporals		
	ARGA	External Causer Argument		
Link Arguments	LINK-PRO	null instantiation of pronoun		
	LINK-PCR	null instantiation of pragmatic coreference		
	LINK-SLC	null instantiation of selection constraint link		

Table 3.2 SR Labels in Two SR Labellers (English and Chinese)

of-art SR labeller PathLSTM (Roth and Lapata, 2016) for English and the Language Technology Platform API (Che et al., 2010), an integrated Chinese processing platform for Chinese SR labelling have mismatches between their SR labels. Table 3.2 shows the differences between two systems. This discrepancy is problematic for comparing the two languages.

In addition, the distribution of SR labels are rather unbalanced, with the first three numbered arguments (A0-3) accounting for around 70% of all types (Hajič et al., 2009). The majority of SR classes do not occur so frequently as the first three. In particular, those function modifiers are comparatively much rarer. In translation, these adjuncts are so free in forms (because of the different creativity or competence of individual translator) that even some of them, for instance, a temporal modifier, may not occur at all in the translation. It is thus viable to group them into fewer categories for the sake of avoiding data sparsity. As a result, I propose to regroup the semantic labels into four major groups, as is shown in Table 3.3.

In their study, Giménez and Màrquez (2007) show that metrics based on the syntactic and shallow-semantic information are able to produce more reliable system rankings than those lexical-oriented metrics (such as BLEU, NIST). As is shown in their experiments, at the shallow semantic level, SR-related similarity metrics (with reference translations) proved very effective and are among the top-scoring in both single-reference and multiple-reference evaluation scenarios. Their findings

SR Labels	English	Chinese
	English (EPB)	Chinese (CPB)
Arg0	agent, experiencer	
Arg1	patient, theme	
Arg2	benefactive/instrument/attribute/end state	
Others	start point/benefactive/instrument/attribute & function modifiers	

Table 3.3 SRL Tags for English & Chinese

suggest that it is important to translate lexical items according to the semantic role they play inside the sentence. However, as Giménez and Màrquez (2007) posit, SR similarity metric, focusing only on partial aspects of quality, does not provide a global measure of quality. They argue that other similarity metrics at different linguistic levels, such as dependency parsing, constituency parsing, should also be integrated into a single measure. In the case of the HTQE task, this implies SR can also be part of the large feature set, especially when NLP techniques have advanced to recognize SRs with good accuracy[8] the semantic roles of arguments in a sentence.

My approach differs from MTQE-related research in which SR similarity is used as a direct metric. Instead, I integrate the normalised frequency information (as in Equation (3.1)) of regrouped SR labels into a broad range of quality indicators, and SR information in both STs and TTs are considered. In addition, the proposed method of regrouping allows the comparability of SRs in different languages for QE in particular.

3.2.1.5 Discourse Aware Features

It is desirable for any text generation to produce coherent texts. For instance, with regard to human translation, we expect smooth grammatical and lexical relations between words and sentences. In other words, the produced translation should be structurally and meaningfully established as a whole, particularly at the document level. It is insufficient to sequentially translate sentences of a source text and then concatenate them in order to obtain the final output. From a linguistic point of view, a high-quality translation should take into account the discourse-wide context (Hardmeier, 2014; Hatim and Mason, 2014). Therefore, I come up with features measuring specifically the **cohesion** and **coherence** of the target text.

In MT, explicitly discourse-related research topics became popular in the research community. Refer to Wong and Kit (2012) for a better review. The idea of translating at the document level and taking into account broader contextual information is to obtain adequate translations respecting cross-sentence relations,

[8] on the OntoNotes benchmark, Peters et al. (2018) has achieved the state-of-the-art 84.6 F-score for English and on the standard benchmark dataset CPB 1.0 (https://catalog.ldc.upenn.edu/LDC2005T23), Sha et al. (2016) has achieved 77.69 for Chinese semantic role labelling.

enforcing cohesion and consistency at the document level (Ben et al., 2013; Xiong et al., 2015). There have been some efforts to exploit discourse information to improve the evaluation of MT in general (Hardmeier and Federico, 2010; Meyer et al., 2012), as evidenced by the biannual Workshop on Discourse in Machine Translation (DiscoMT) (Webber et al., 2013, 2015, 2017). Engineering discourse-related features to widen the scope of QE is one of the common strategies to the document-level translation evaluation. Many metrics in these evaluation campaigns and quality estimation tasks explore the ways to incorporate semantic, syntactic and discourse features. Comelles et al. (2010) design and extend a set of discourse representation features, e.g. lexical overlap between discourse representation structures of the same type (Giménez and Màrquez, 2009), syntactic tree matching (Liu and Gildea, 2005), to evaluate document-level newswire MT translations. By now, several discourse-focused research problems have been actively explored in MT, such as predicting a target-language pronoun given a source-language pronoun in the context of a sentence and/or a full document, inter-and interlingual variation of discourse phenomena (Lapshinova-Koltunski, 2015), coreference resolution (Novák et al., 2015), lexical consistence (Guillou, 2013), discourse connectives (Meyer and Webber, 2013; Steele, 2015).

Crucial to the measurement of cohesion is the differentiation between cohesion and coherence. According to Brunette (2000), coherence can be defined as the 'continuity of the meaning of a text from one idea to another and plausibility of such meaning', and cohesion the 'linguistic means used to ensure continuity of the form and content of a text'. The author claims that checking whether the translation is sufficiently well linked on a semantic coherence and formal cohesion level to constitute an effective text for the target language community often makes the first step of quality assessment. Some discourse-related indexes have been proposed to measure MT quality. Giménez et al. (2010) presented a modified MT evaluation metric based on Discourse Representation Theory (Kamp and Reyle, 2013), which employs features based on coreference relations and discourse connection to assess the quality of MT output. Wong and Kit (2012) proposed to use word repetition to measure the lexical cohesion in texts. Scarton et al. (2016), following their previous work (Scarton et al., 2015a), use discourse-related features, such as pronouns, connectives, elementary discourse unit[9](EDU) (Mann and Thompson, 1988), in addition to latent semantic analysis cohesion features (e.g. average LSA correlation of adjacent sentences, or of all sentences). In the similar vein, Joty et al. (2017) use the sentence-level discourse structure based on Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) trees and sub-tree kernels. However, it is worth noting that RST tree features heavily rely on external discourse parsing tools that could only work with the English STs or TTs, and there are no readily available resources to train an RST parser for Chinese. Most importantly, the RST

[9]the minimal building blocks of a discourse tree, controversially consisting of sentences, prosodic units, turns of talk, clauses etc.

parsing itself is essentially constricted to the sentence level, which does not fully measure the textual coherence.

Thus, in this study, I measure the quality of translations with some **explicit cohesion devices and implicit coherence** indexes.

- **Cohesion Features**

  Unlike MT working mainly at the sentence level, human translators rarely consider only the isolated sentences or sub-sentential segments. The coherence between the ST and the TT and within the TT itself and the cohesion in the ST impact a translator's translation process and final performance as well. Halliday and Hasan (2014) identified references, substitution, ellipsis, conjunction and lexical cohesion as five main categories of cohesion in English. The first 4 are roughly grammatical cohesion in contrast to lexical cohesion that connects sentences through lexical choices. Inspired by the recent work of Crossley et al. (2016a,b) on cohesion in writing, I create a set of cohesion features for HTQE, which are specifically adapted for capturing grammatical cohesion for the purpose of translation evaluation. For cohesion features, I focus on features of repetition, pronoun reference and linking connectives. For connectives, I compile a bilingual lexicon of discourse connectives based on the English list provided in Crossley et al. (2016b). Note that unlike the oft-quoted theoretically-based connectives in Halliday and Hasan (2014) that are divided in five categories: causal (because, so), contrastive (although, whereas), additive (moreover, and), logical (or, and) and temporal (first, until), the framework of connective indices used is based on rhetorical features. To this end, I query a bilingual English-Chinese online dictionary[10] and manually analyse the concordance lines of the parallel corpus[11], linking the Chinese translations to those English connectives. Table 3.4 lists instances of ST and TT sentences that contain the corresponding connectives. Through some manual checking of randomly selected connectives, the researcher confirms that the bilingual lexicon of connectives covers the majority of correctly translated connectors in the corpus of trainee translations[12].

  - **argument type/token ratio**
    the number of unique nouns and pronouns divided by the number of total nouns and pronouns (in tokens)

  - **pronoun density**
    the number of pronouns divided by the number of words

[10] http://dict.youdao.com/, which is a mega-size dictionary, including Longman, Collins, New English-Chinese etc.

[11] https://www.linguee.com/

[12] The researcher randomly queries the translations for 10 English connectives in the ST and find 9 of them are in the list of the bilingual lexicon. The only mismatch is caused by orthography. The student translated 'then' as 于似乎 (于是乎).

- **pronoun-noun ratio**
  the number of pronouns divided by the number of nouns
- **pronoun-noun phrase ratio**
  the number of pronouns divided by the number of noun phrases
- **content word repetition**
  the number of the repeated content words divided by the number of words
- **adjacent sentence overlapping**
  overlapping words of any two adjacent sentences divided by the total number of words in two sentences
- **basic connective**
  the number of basic connectives divided by the number of words in the text
- **simple subordinators**
  the number of subordinators divided by the number of words in the text
- **coordinating conjuncts**
  the number of coordinating conjuncts divided by the number of words in the text
- **addition**
  the number of addition words divided by the number of words in the text
- **sentence linking**
  the number of sentence linking words divided by the number of words in the text
- **order**
  the number of order words divided by the number of words in the text
- **reason and purpose connective**
  the number of reason and purpose connectives divided by the number of words in the text
- **demonstratives**
  the number of demonstratives divided by the number of words in the text
- **additive**
  the number of additive connectives divided by the number of words in the text
- **determiners**
  the number of determiners divided by the number of words in the text

– **causal**
the number of casual connectives divided by the number of words in the text

– **logical**
the number of logical connectives divided by the number of words in the text.

– **semantic cohesion**
averaged LSA similarity between all sentences.

ST	TT	Connective Types
The partners may have some interests in common , but these interests are generally insignificant .	他们 彼此之间 会 有 一些 共同 的 爱好 ， 但是 这些 共同 的 爱好 一般 都 是 没有 多大 意义 的 。	Basic
"Imagine , " I said , "if I 'd had a vision and worked at it , just a little bit every day , what might have I accomplished ? "	"我 说 ： "" 想像 一下 , 如果 我 有 一个 打算 并且 致力于 完成 它 , 每天 哪怕 仅为 之 做 一点 , 那 我 现在 将会 有 多 大 的 成就 啊 ? ""	subordinator
They may also quarrel in public or put up a facade of being compatible .	他们 也 可能 在 公共场合 吵架 ， 或者 是 表现 在表面上 。	addition
"however , the science of xenotransplantation is much less straightforward . "	然而 ， 异种 器官移植 的 科学 可不是 那么 容易 探索 的 。	linkings
I first took up walking as a means of escape .	"我 最初 开始 散步 是 为了 要 逃避 一些 事情 。"	order
that he is deeply concerned with the charm and quality of things , and gentlest light , so that at least he may make others love life a little better ,	他 被 事物 的 魅力 和 品质 所 吸引 ， 所以 至少 他 能 使 其它 人 更加 地 热爱生活 ， 并 使 他们 准备 好来 迎接 生活 的 多种多样	Reason-purpose

Continued on next page

Table – *Continued from previous page*

ST	TT	Connective Types
"He does not see life as the historian , or as the philosopher , or as the poet , or as the novelist , and yet he has a touch of all these . "	他 不以 历史学家 的 视角 来 看待 生活 ， 也 不以 哲学家 、 诗人 或是 小说家 的 身份 来看 ， 然而 他 的 思想 涉猎 到 这些 所有 职业 。	opposition
"The couples in these marriages engage in few activities together and display no pleasure in being in one another 's company . "	这种 婚姻 中 的 夫妻 很少 在 公共场合 共同 露面 并且 对 对方 的 工作 显示 不出 任何 乐趣 所在 。	demonstrative
During 1996 at least two big reports on the subject – one in europe and one in america – were published .	1996 年间 ， 至少 两起 大 的 异种 移植 的 报道 被 发表 ， 一起 在 欧洲 ， 一起 在 美洲 。	logical

Table 3.4 Example of Bilingual Lexicon of Connectives

- **Coherence Features**

  Morris and Hirst (1991) posit that text or discourse is a set of sentences that tend to be about the same things, i.e. having a quality of unity, which is a property of coherence. The sameness of text can be achieved by 'preserving the relatedness of the group of words' (Klebanov and Flor, 2013). Xiong and Zhang (2013) propose a topic-based coherence model to predict the target coherence chain with the extracted source coherence chain by which a document can be represented as a continuous change of topics. Ben et al. (2013) propose a bilingual lexical cohesion trigger model to model the co-occurrence of the source language lexical cohesion item and its target language counterpart, using mutual information to measure the strength of their dependency. Following the same line of thought, Klebanov and Flor (2013) first build word association profile, i.e. pointwise mutual information (PMI), for all pairs of content words from a very large and diverse corpus, and then compute the average of word association profile (termed as **lexical tightness**) for each target translation. They have shown that translated texts are less lexically tight than the originals and that better translations are tighter than worse translation in terms of average value of PMIs in the text under evaluation. In line with these studies, I compute the lexical tightness of each translation, with a slightly more complex method of building the word association profile. In Klebanov and Flor's work (2013), they compute PMIs using all co-occurrence

counts of content words in the same paragraph from a large and diverse corpus. However, I argue that content words do not necessarily collocate with other content words beyond a sentence boundary. Thus, there is no point in building an association profile for words which are separated too far from each other in the paragraph. For instance,

> The life of a worker honeybee is even separated into successive occupations : during the first three weeks the young worker grooms the queen and her eggs , cleans out the hive , cools it by wing-fanning at the entrance , and attacks or walls in intruders . Only after this apprenticeship is the graduate allowed to leave the hive and forage for nectar and pollen . Add to such behaviour the fact that some ants use leaf fragments as spoons in which to carry soft food back to their nest , and one is tempted to describe insects as " intelligent " and begin to make comparisons between insect and human societies .

in which 'life' in the first sentence and 'hive' in the second sentence will co-occur less frequently than 'allow' and 'leave' in the second sentence. Therefore, I apply a sliding window strategy to the content words in the same paragraph where I compute PMIs for words appearing within a $[-5, 5]$ (left 5 words and right 5 words) from the Chinese Wikipedia Dump[13] for all translations. This technique significantly reduces the size of the association profile, speeding up the training process. To represent the translated sentences and documents, I average all the PMIs according to the PMI association profile.

Other than the lexical tightness of the text, I also measure the string-based and word-representation-based distance and correlation

- **lexical tightness**
  Pointwise Mutual Informaiton (PMI) (Church and Hanks, 1990) of content words in a text

- **averaged cosine distances of adjacent sentences**
  computed from vectors of bag-of-words.

  Textual data are converted into d-dimensional vectors of numbers reflect various linguistic properties of them, and the bag-of-words approach looks at the histogram of the words within the text, considering each word count as a feature (Goldberg, 2017). The averaged cosine distances of adjacent sentences in a document are used to represent the content coherence within the document. Cosine distance is obtained as

$$\mathrm{D_c}(\mathrm{A}, \mathrm{B}) = 1 - \mathrm{S_c}(\mathrm{A}, \mathrm{B}),$$

[13] https://dumps.wikimedia.org/zhwiki/20171103/zhwiki-20171103-pages-articles-multistream.xml.bz2. Detailed description of the corpus is provided in Section 5.3.1

ID	Definition	# Features
F1-24	PoS tags [S & T]	2*12
F25-98△	Dependencies [S & T]	2*37
F99-112	Shallow Features [S & T] 2*7	
F113-124	Constituency [S & T]	2*6
F125-132△	Semantic role labels [S & T]	2*4
F133△	Constituency parsing probability [T]	1
F134-174△	Cohesion and coherence features [S & T]	41

Table 3.5 Monolingual Features

where D_c denotes the cosine distance and S_c is the cosine similarity that can be computed by Equation 3.3.

– **averaged cosine distances of adjacent sentences**
computed from vectors of word embeddings.

Recent trends suggest that neural network-inspired word embedding models outperform traditional count-based distributional models in many NLP tasks such as POS tagging (Collobert and Weston, 2008; Collobert et al., 2011), analogy detection (Socher et al., 2012; Mikolov et al., 2013a), sentiment analysis (Socher et al., 2011; Dickinson and Hu, 2015), textual entailment (Bjerva et al., 2014; Zhao et al., 2015). As Li and Yang (2018) state, one of the most common applications of word embeddings is semantic analysis, in which nearly half of them involve word embedding.

In this work, adjacent sentence cosine distance is calculated by computing the mean of the cosine distances of the concatenated vectors of words (summation) in adjacent sentences.

– **averaged correlation distance of adjacent sentences**
computed from vectors of bag-of-words.

– **averaged correlation distance of adjacent sentences**
computed from vectors of word embeddings

3.2.1.6 Other Shallow Features

In addition to the features listed above, a group of shallow statistics of the ST and the TT are included. They mainly consist of:

- **Types**
that refer to the size of the lexicon, i.e. a set of unique words, in running text.

- **Tokens**
that refer to all words in running text, i.e. the text size.

- **Type-Token Ratio, TTR**
  that stands for the proportion of the number of types to the number of tokens in the running text. TTR is normalised as per Equation 3.1.
- **Number of Sentences**
- **Averaged Sentence Length**
- **Number of Content Words**
  which are lexical items having a relatively 'specific or detailed' semantic content (Corver and Riemsdijk, 2001), including nouns, verbs, adjectives and adverbs.
- **Number of Function Words**
  which carry little lexical meaning and express grammatical relationships among other words within a sentence. For instance, function words in English include determiners, conjunctions, prepositions, pronouns, auxiliary verbs, modals, qualifiers and question words.

Thus far, the researcher has introduced the main components of this group of features. Under the monolingual category, a detailed list of monolingual features in both ST and TT is shown in Table 3.5 (triangle marked items are newly designed features). Note that these features are computed at the document-level, and in the case of sentence level QE, the adjustment in calculation should be accordingly made. For instance, treating each sentence as an individual document and computing the same set of features on its basis. In case such treatment that does not apply, I remove the corresponding feature (s).

3.2.2 Bilingual Features

Bilingual features are linguistic variables linking STs and TTs in a dynamic way. As a common practice of human translation, inspecting whether core information and the ST features are completely and successfully delivered in the TT is a way of measuring translation quality. For QE, I am looking for ways of simulating human beings' scrutiny of translation quality, i.e. comparing ST and TT from the perspectives of all sorts.

3.2.2.1 Log Ratios of Paired Monolingual Features

Features that closely link ST and TT should be taken into account. In this regard, The logarithmized ratio is adopted to make manageable the ratios of features pairs coexisting in STs and TTs. For instance, the log ratio of ST and TT tokens can be a feature to measure the completeness of the translation. I compute the logarithmized ratios by taking the base 10 logarithm of the ratio of each feature between TT and ST. To avoid the problem of zero division, I add a small floating number $\mathrm{f}(0.001)$ to

both the numerator and the denominator. Equation (3.2) shows how such features are calculated.

$$\mathrm{F}_{\mathrm{log\_ratio}} = \log_{10} \frac{\mathrm{Count}_{\mathrm{tt}} + \mathrm{f}_{[0.001]}}{\mathrm{Count}_{\mathrm{st}} + \mathrm{f}_{[0.001]}} \tag{3.2}$$

3.2.2.2 ST-TT Distance

City Block Distance (CBD) is also known as Manhattan distance. It measures the distance between two points in a Euclidean space. On the assumption that the closer two points are the more similar they are to each other. Smaller CBD values, therefore, indicate better translation quality. In this study, STs and TTs are treated as two Euclidean points in a space, and their distance is measured by CBD. This concept is borrowed to quantify the distance between an ST and a TT in terms of a specific feature pair or group. For example, the ST and TT token-type CBD refers to the ST and TT distance between the two vectors of two elements (counts of types and token in both directions). CBD can be computed as

$$\mathrm{d}_{(\mathrm{ij})} = \sum_{\mathrm{i}=1}^{\mathrm{n}} |\mathrm{x}_{\mathrm{ik}} - \mathrm{x}_{\mathrm{jk}}|,$$

where x_{ik} and x_{jk} are both k dimension vectors (i.e. the size of coordinates in them is k). Take the Semantic roles in both ST and TT as an example, as they are divided into four groups, i.e. A0, A1, A2, ARGM, the CBD of semantic role labels between ST and TT then should be two 1×4 vectors. Therefore, if the normalised frequencies of the four SR labels are $[7, 26, 4, 11]$ and the corresponding counts of SR labels in TT is $[10, 33, 5, 26]$, the ST and TT SR CBD should be the absolute differences of 4 pairs of coordinates, as is shown below.

$$\begin{aligned} \mathrm{D}_{\mathrm{SR}} &= |7-10| + |26-33| + |4-5| + |11-26| \\ &= (3+7+1+15) \\ &= 26 \end{aligned}$$

This CBD value 26, calculated from the summation of the differences in SR label frequencies, can be part of the quality-indicative features.

3.2.2.3 Pseudo-reference Agreement Scores

Pseudo-references are substandard reference translations generated by other MT systems. As the name suggests, they are not classic ‘gold standards’ and can be varying in quality themselves, not necessarily as good as or better than the target translation under evaluation. The main theory of this approach is that through regression learning the trained metric (a learned function) is able to map a feature vector (signifying target translation’s similarity to pseudo references) to a quality score of whatever scheme. In this sense, imperfect translations are also informative, despite

that we do not know the actual distance and the quality of the pseudo reference (s). The concept of pseudo-reference was proposed to capture the adequacy of translations and address the issues of combining evidence from heuristic distances and calibrating the quality of pseudo reference system (Albrecht and Hwa, 2007b). I will test this working assumption with our data in Chapter 5

In MTQE, pseudo-references are used as references to compare with target MT translations using the well-known metrics, such as BLEU, HTER, METEOR, TER (Snover et al., 2006). Researchers tried to integrate scores from the aforementioned metrics to their feature set for QE. While the model of pseudo-references treats the metric as a distance measure without human references, features are adapted from the common standard distance metrics, such as BLEU, ROUGE (Lin, 2004), METEOR, Head Word Chain (HWC) (Liu and Gildea, 2005). To a certain extent, these studies have shown that metrics developed using regression learning based on a set of pseudo-references rival standard reference-based metrics in terms of correlations with human judgements. Pseudo-references are found to be informative comparison points. Metrics trained this way often have a higher correlation with human judgements than standard metrics based on multiple human references. Most importantly, better target translations can even be predicted by the worse pseudo references (Albrecht and Hwa, 2007a,b; Scarton and Specia, 2014a; Langlois, 2015).

To follow this approach, I use three MT systems (Google Translate[14], Bing Translator[15] and Yandex Translate [16]) to generate translations as references and then calculate the 'similarity' between pseudo-references and translations as the pseudo-reference agreement scores. To this end, a set of new features are proposed, which include:

- **TFIDF Cosine Similarity**
  Each target translation and the pseudo references are treated as a bag of words and n-grams (1-3 in this study), and they are converted to a matrix of sparse TF-IDF features, which are calculated in the form:

$$\text{tf-idf}_{t,d} = (1 + \log \text{tf}_{t,d}) \cdot \log \frac{N}{\text{df}_t}.$$

  Cosine similarity, which is defined as in Equation 3.3, is then obtained to represent the distance between the target translation and the paired reference translation. It is worth pointing out that I choose different similarity and distance metrics in order to increase the diversity of features representations of STs and TTs. The inclusion of other new metrics is mostly out of the same motivation. Measuring the orientation of two n-dimensional vectors, the cosine similarity computes document similarity (Xu and Wunsch, 2005). It calculates the dot product of two numeric vectors, normalised by the product of the vector

[14] https://translate.google.co.uk/
[15] https://www.bing.com/translator
[16] http://https://translate.yandex.com/

lengths. Intuitively, the higher its cosine similarity score is, the more likely two translations match each other literally. Thus, their quality should be close. I use cosine similarity for comparing the two vectors for both the target translation and the reference. It is calculated as

$$\begin{aligned} \cos(\mathbf{x}, \mathbf{y}) &= \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|} \\ &= \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} {x_i}^2} \cdot \sqrt{\sum_{i=1}^{n} {y_i}^2}}, \end{aligned} \tag{3.3}$$

where x_i and y_i denote elements of vector $\mathbf{x}$ and $\mathbf{y}$ respectively. The resulting similarity is bounded between 0 meaning irrelevant relationship and 1 suggesting exactly the same. In our case, the elements of vectors are word frequencies of the sentences or documents. For each translation, I calculate a cosine similarity score based on translation of each system and I also compute a geometric mean of all three similarities. Therefore, I obtain 4 MT pseudo-translation similarity scores for each translation.

- Geometric Mean of MT Cosine Similarity
  To minimise the over-influence of any specific pseudo-reference similarity in measuring the quality of the target translation, I compute the geometric mean of the cosine similarity between the input and three pseudo system outputs.

  In mathematics, the geometric mean is one of the three classical Pythagorean means, indicating the central tendency of a collection of observations. The other two are the arithmetic mean (i.e. the sum of a set of numbers divided by the number of the numbers in the set) and the harmonic mean (i.e. the reciprocal of the arithmetic mean of the the reciprocals of the collection of numbers). A geometric mean is defined as the nth root of the product of n numbers (Mohanty and Kumar, 2015, pp.33):

$$\mathrm{GM} = \left(\prod_{i=1}^{N} x_i \right) = \sqrt[n]{x_1 x_2 \dots x_n}.$$

- **Levenshtein Distance**
  The Levenshtein distance is a numerical representation of the minimum cost of insertions, deletions or substitutions that transform one string into another. It compares sequences, such as phone string, dialect pronunciation, text similarity, gene sequence (Heeringa, 2004). In respect of target translations and pseudo-references, the running texts or sentences can be viewed as a whole sequence of characters or letters and compared. Pseudo-references and translations that have smaller Levenshtein values are likely to be overlapping in content.

Fundamental to the idea of the Levenshtein distance is the notion of string changing operations, a collection of deletions, substitutions and insertions that determine the extent to which two strings differ from each other. Levenshtein distance can be normalised in various ways (Marzal and Vidal, 1993; Weigel and Fein, 1994; Li and Liu, 2007). Among various approaches to normalisation, I select the normalisation based on the shortest and longest alignment length (Heeringa, 2004) because of its simplicity and easy implementation, in addition to its taking into account the alignment nature of translation as a regularisation means. Thus, I have three metrics based on Levenshtein distances, i.e. the original Levenshtein, the normalised one based on the short alignment and the normalised one based on the longest alignment, for measuring the string similarity between the pseudo-references and the target translation.

- **Sorensen Distance**
  The Sorensen distance, also known as Dice's coefficient, is a metric originally designed for the comparison of biological specimens, and it yields a real numeric value between 0 and 1. This metric can be employed to describe the lexical similarity between two sequences of strings (Thomas and Short, 1996, pp.217). The formula for Sorensen distance is then given as:

$$\mathrm{QS} = \frac{2 \mid \mathrm{X} \cap \mathrm{Y} \mid}{\mid \mathrm{X} \mid + \mid \mathrm{Y} \mid},$$

  where $\mid \mathrm{X} \mid$ and $\mid \mathrm{Y} \mid$ denote the number of substrings (e.g. bigrams) in the two sequences, and $\mid \mathrm{X} \cap \mathrm{Y} \mid$ represents the number of matching bigrams.

- Jaccard Distance
  The Jaccard distance is complementary to the well-known Jaccard similarity coefficient for measuring the relative size of the overlap of two finite sets A and B (Kosub, 2016). It is defined as:

$$\mathrm{d_j}(\mathrm{A}, \mathrm{B}) = 1 - \mathrm{J}(\mathrm{A}, \mathrm{B}) = \frac{\mid \mathrm{A} \cup \mathrm{B} \mid - \mid \mathrm{A} \cap \mathrm{B} \mid}{\mid \mathrm{A} \cup \mathrm{B} \mid},$$

  where $\mid \mathrm{X} \cup \mathrm{Y} \mid$ is the size of union and $\mid \mathrm{X} \cap \mathrm{Y} \mid$ is the intersection of two sets. Therefore, the metric is equivalent to subtract the Jaccard coefficient from 1. This metric is, among many others, one of the classical text similarity measures to compare words, sentences, paragraphs and documents from a string-based perspective (Gomaa and Fahmy, 2013).

- **TFIDF Weighted Word Representation Similarity**

  In this work, I implement word embedding based sentence similarity calculation by computing cosine distance of the mean of the concatenated vectors of words in a sentence or a document in the target translation and its corresponding

mean of the concatenated vectors of words in the sentence or document of the pseudo-references. While it is viable to compute the cosine distance of the averaged vectors that are concatenated after removing the stopping words in a sentence or a document, I propose to use the mean of the TF-IDF weighted vectors for both target translations and pseudo-references in order to give more weight to those more meaningful words in a sentence or document.

- **Similarity-based MT Evaluation Metrics**

  Though pseudo-references are not human translations, features adapted from the above-mentioned distance-based metrics can be input to the selected learning algorithm and combined mathematically with other features to quantify the target translation's quality. I choose the following MTQE metrics because they are established standards representative of different focuses while MT translations are evaluated. Another reason is that they are already included in the MTEval toolkit[17], which saves us from the trouble of reimplementing them from scratch.

  - BLEU
  - RIBES
  - NIST
  - WER

  Take BLEU as an example. To measure the distance between the target translation and pseudo-translation, I obtain the sentence-level BLEU score and corpus-level BLEU score. The only difference between the two scores is that corpus-level BLEU score involves comparing the target translation with more than one pseudo-reference, while sentence-level BLEU score means comparing one target translation with only one pseudo-reference. In either case, the whole documents of target translations, when evaluated as a whole, are treated as individual sentences. This helps avoid the problem of incomplete target translations (students sometimes produce shorter, misaligned translations), which makes it impractical to compare them with pseudo-reference pairwise at the sentence-level (i.e. one translation sentence v.s. one pseudo-reference sentence) and the corpus level (i.e. one translation sentence v.s. multiple pseudo-reference sentences). I compute the BLEU score for each target translation against each pseudo-reference (sentence-level BLEU) and against them all (corpus-level BLEU).

  Note that RIBES has not been introduced in Section 2.2. It is short for rank-based intuitive bilingual evaluation score. The metric is designed to tackle

[17] https://github.com/odashi/mteval

the problem that word order is not significantly penalised in conventional metrics, an element particularly important for translation quality within distant language pairs, such as English-Chinese and English-Japanese (Isozaki et al., 2010). RIBES is based on rank correlation coefficients that compare the word ranks in the reference and the target translation. Normalised Kendall τ (Kendall, 1938) and Spearman ρ are used to compute the word rank correlation, penalised by the square root of precision (the number of corresponding words in proportion to the number of words in the target translation). Spearman correlation coefficient ρ and Kendall τ are defined as in Equation 3.4 and Equation 3.5.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{3.4}$$

where n is the number of smaples and d is the pairwise distances of the ranks of the variables x_i and y_i.

$$\tau = \frac{S}{\sqrt{n(n-1)/2 - T}\sqrt{n(n-1)/2 - U}} \tag{3.5a}$$

$$T = \sum_t t(t-1)/2 \tag{3.5b}$$

$$U = \sum_u u(u-1)/2 \tag{3.5c}$$

where S is the difference between the number of concordant pairs and the number of discordant pairs, t is the number of observations of variable x that are tied and u is the number of observations of variable y that are tied. These rank correlation metrics are then normalised to be within the range $[0, 1]$ as follows.

$$\hat{\tau} = \frac{(\tau + 1)}{2}$$
$$\hat{\rho} = \frac{(\rho + 1)}{2}$$

However, the metrics will overestimate incomplete translations if those few words contained in the translation appear in the order as they are in the reference. Therefore, the common metrics precision, recall and F-measure are used to address the overestimation issue. Formally, they are given as

$$P = \frac{c}{|h|},$$

$$R = \frac{c}{r},$$

and

$$F_\beta = \frac{(1+\beta^2)PR}{(\beta^2 P + R)}.$$

Among them, c is the number of corresponding words and $|h|$ is the number of words in the candidate translation, $|r|$ is the number of words in the reference translation and β (default: 0.1) is a parameter. The precision is found to correlate well with adequacy such that I have two new metrics:

$$\hat{\rho}P^\alpha$$
$$\hat{\tau}P^\alpha,$$

with α (default: 0.25) being a parameter in the range $[0, 1]$.

RIBES can be useful to measure translation quality for distant language pairs by giving more prominence to the word ordering in the translations. However, the parametrisation of certain variables such as α in the calculation is less persuasive.

3.2.2.4 MT Back-translation Similarity

Back-translation is 'a special case of the mapping of an equivalent set of sentences in one language onto a set in another' to check translation quality (Brislin, 1995, pp. 32). In back-translation, a second bilingual or translator, who has not seen the ST, translates the TT back into the SL. The purpose of such a process is to obtain a literal version of TT in order to evaluate its semantic equivalence to the ST. This method enables a monolingual of the SL to compare the two SL texts (the original ST and its back-translated version) and make an indirect judgement about the quality of the translation.

Therefore, the working sequence of back-translation would be 'ST-TT-ST (back-translated)-assessment'. To implement back-translation for hundreds and even thousands of (sentences or documents of) human translations, we have to rely on bilingual experts or professional translators to translate the TTs and evaluate their back-translations. This is definitely unaffordable and impractical for translation evaluation in large volumes. Therefore, I come up with an alternative solution: Using three state-of-the-art commercial MT systems (Google Translate, Bing Translator and Yandex Translate) to translate the human translations back to the SL, and to compare the machine back-translated ST with the original ST I compute the similarity scores between them.

For the set of features with respect to back translations, I compute the same set of features as for pseudo-references.

3.2.2.5 Alignment Features

In this section, I present my features of the word and phrase alignments.

- **Word Alignment Features**

  As far as we know, there are three common approaches that are used to find mappings between individual token links in bilingual sentences, namely the **heuristic approach** (word alignment based on co-occurrence) (Melamed, 1995; Moore, 2005), the **stochastic approach** (generative modelling) (Brown et al., 1993) and the **discriminative approach** (supervised or unsupervised learning based) (Liu et al., 2005; Moore et al., 2006; Liu et al., 2009). Most current work on word alignment is generative-modelling, which has been prevalent in SMT. Generative models try to 'build a stochastic model' that can 'translate arbitrary sentences from one language to another', modelling the translation process as a search for the most probabilistic candidate (Tiedemann, 2011, pp.60). Such stochastic models in the framework of SMT often use a latent alignment variable to determine the correspondence of the source words and the target words from parallel sentences using expectation maximisation (EM) (Dempster et al., 1977; Liu et al., 2009). EM algorithm is an efficient iterative procedure to compute the maximum likelihood estimate (MLE) of model parameters for which the observed data are the most likely. MLE is a method that determines values for the parameters of a model such that they maximise the likelihood that the process described by the model could actually be observed (Pfanzagl, 1994, pp. 207). The EM iteration consists of two processes: expectation and maximisation. In the expectation, the probability of the missing data is estimated based on the observation of historical data and the current model parameters, using conditional expectation. In maximisation, the likelihood function is maximised under the assumption that the missing data are known. Figure 3.3 illustrates word alignments between a source English sentence and its parallel Chinese sentence.

  As the figure of alignment shows, an equivalent translation often assumes a large proportion of aligned words ($7/11$) between the ST and TT. This implies that translation quality is to some extent related to how well two sentences are aligned[18]. In other words, the more aligned words we can find between two sentences, the more likely they are translations of each other, given that both sentences read grammatically and naturally.

  Recent years have seen attempts using word alignment information for MTQE (Ueffing et al., 2003; Bach et al., 2011; Popović et al., 2011; Popovic, 2012; Camargo de Souza et al., 2013; Specia et al., 2015; Abdelsalam et al., 2016; Yuan et al., 2016). As Abdelsalam et al. (2016) noted, the majority of these

[18] It is indeed the case with literal translations which use very light paraphrasing.

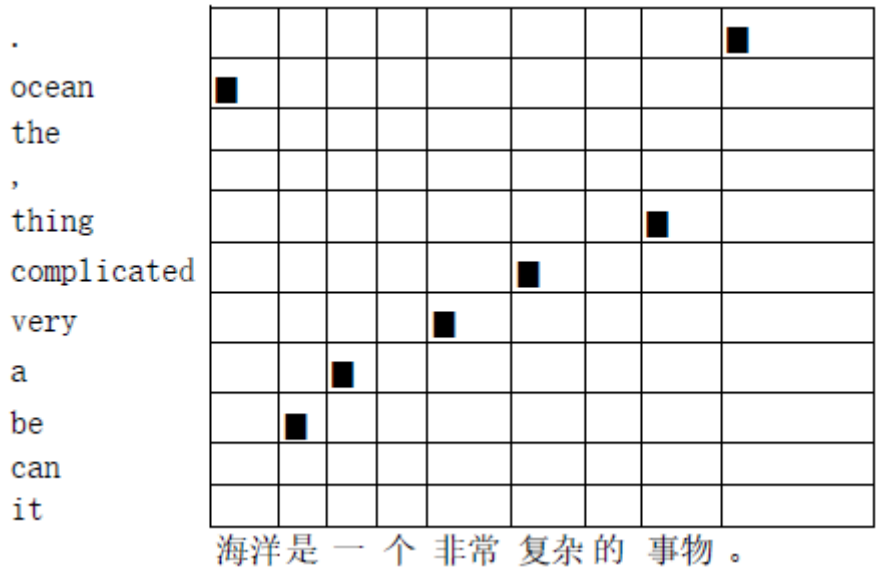


Figure 3.3 An Example of Word Alignment

research focuses on exploiting alignment related information for word-level QE. Among the few studies (Camargo de Souza et al., 2013; Abdelsalam et al., 2016; Yuan et al., 2016) actually try to tackle QE at the sentence-level or above, some features are too complex to be interpretable to humans. For instance, Bach et al. (2011) use the source and target alignment context and even combine the alignment context with POS tags, and Camargo de Souza et al. (2013) implement features, such as proportion of alignments connecting words with the same POS tag and proportion of words in ST and TT that share the same POS tag. I argue that on the one hand, such features increase feature complexity, and on the other hand, they are against the intuition that linguistic attributes, such as POS, may not remain the same during the translation for two drastically different language pairs. I continue my way of obtaining alignment precision and recall in Yuan et al. (2016), which computes the proportion of aligned words in source sentences or documents (precision) and the proportion of aligned words in target sentences or documents (recall). They are similar to two of many alignment features[19] used by Camargo de Souza et al. (2013). However, my word alignment features differ by considering the sentence or document length information of ST and TT, and they are normalised. I also propose that the summation of the logarithmized probabilities (IBM scores) of all aligned words in the documents (sentences) could be a potential quality indicator.

Therefore, we have three-word alignment features:

- normalised proportion of aligned words by ST length (word alignment precision)

[19]They use proportion of aligned words and proportion of aligned n-grams. The latter is similar to the proposed feature of phrase alignments.

  - normalised proportion of aligned words by TT length (word alignment recall)
  - summation of the logarithmic probability scores of aligned words

- **Phrase Alignment Features**
  In SMT, bilingual phrasal units in a relaxed definition are useful translational knowledge, which is an essential part of statistical systems. Such a knowledge resource is built from parallel corpora with alignment algorithms in the form of 'phrase table' listing word or collocation translations, translation examples at various granularities, or generalisation of transferring rules and/or patterns for translation (Wu, 2010). In SMT, interest in phrase alignment is more about the methods of alignment (Koehn et al., 2003; Deng and Byrne, 2005; Zhang and Vogel, 2005; DeNero and Klein, 2008; Neubig et al., 2011; Hewavitharana and Vogel, 2016) to improve system performance and the application of phrase-based models in other NLP scenarios (Imamura, 2002; MacCartney et al., 2008) as well as in MT. However, little research has been done to explore the application of phrase alignment in QE. To the best of my knowledge, this is the first attempt to apply information regarding phrase alignments to HTQE. I am particularly interested to know how bilingual phrasal units mined from large corpora can be used as a resource for HTQE.

  It is often observed that human translators translate groups of words as a whole and words are rarely treated as the working translation units individually. Translation variation from a large corpus and distribution of frequencies, as illustrated in Table 3.6, will lead to varying probabilities of the aligned phrases. For example, we query 'It is generally accepted that' on an online parallel corpora[20] and manually analyse the returned concordance lines so that we can identify the frequencies for each corresponding translation segment in Chinese. Though these candidate translations may not be exhaustive, students' translations, if not among this list, are likely to be inappropriate translations.

  Therefore, I propose that looking for the consistency of phrasal alignments in trainee translations in relation to professional translations would be a viable indicator of their quality. Under this circumstance, measuring the quality of a translation pair has turned into the query of successful alignments in a database, in which phrase (including word) alignments are learned directly from a large training corpus. In the translation pairs, phrases, including words, are looked up sequentially in the bilingual lexicon that is learned directly from a bilingual corpus through an alignment process. The context and neighbouring words to a sequence or phrase have been taken into consideration by the alignment algorithm. Variations of handling ambiguity in the translations are reflected in the aligned words or phrases themselves. Therefore, more

[20] Generated from http://www.linguee.com

Source	Translation	Gloss	Frequency
It is generally accepted that	一般 认为	generally believe	4
	普遍 接受的 看法	generally accepted opinion	4
	一般 公认	general public acknowledgement	2
	人们 普遍 认为	people generally believe	4
	一般 意见 认为	general ideas believe	2
	一般 看法 是	general opinion is	4
	各国 普遍 赞同	states generally agree	1
	众所 公认 的 是	what public agree is	1
	社会 普遍 接受	societies generally accept	1

Table 3.6 Professional Translations for the Same Source Text Phrase

percentage of phrasal alignments found in a student's translation can be interpreted as a higher degree of semantic adequacy and stylistic fluency. Similar to word alignment features, I design three phrase alignments features for STs and TTs. They are:

- normalised proportion of aligned phrases by ST length (phrase alignment precision)
- normalised proportion of aligned phrases by TT length (phrase alignment recall)
- summation of the logarithmic probability scores of aligned phrases

3.2.2.6 Bilingual Word Representations

Difficulties arise, however, when we try to recognise and quantify semantic similarities across languages. Recognising bilingual semantic similarity is critical for the identification of the equivalence of STs and TTs. After all, translation success places much importance on the adequacy of TTs in relation to the STs. To quantify this semantic adequacy or similarity between texts in two or more languages, we need a method to tackle the non-occurrences of word pairs in training parallel texts and help determine the similarity of bilingual texts.

Word representations (also known as embeddings) provide a viable solution, capable of capturing not only semantic contexts across different languages but also syntactic information (Zou et al., 2013). This special property makes it very useful for quality evaluation. The concept of word embedding was originally put forward by Bengio et al. (2003) as the distributed representation for words, simultaneously representing each word and the probability function for word sequences on the assumption that similar words have similar nearby representations. To illustrate, I plot the first 500 words for better presentation by projecting into a 2D space their

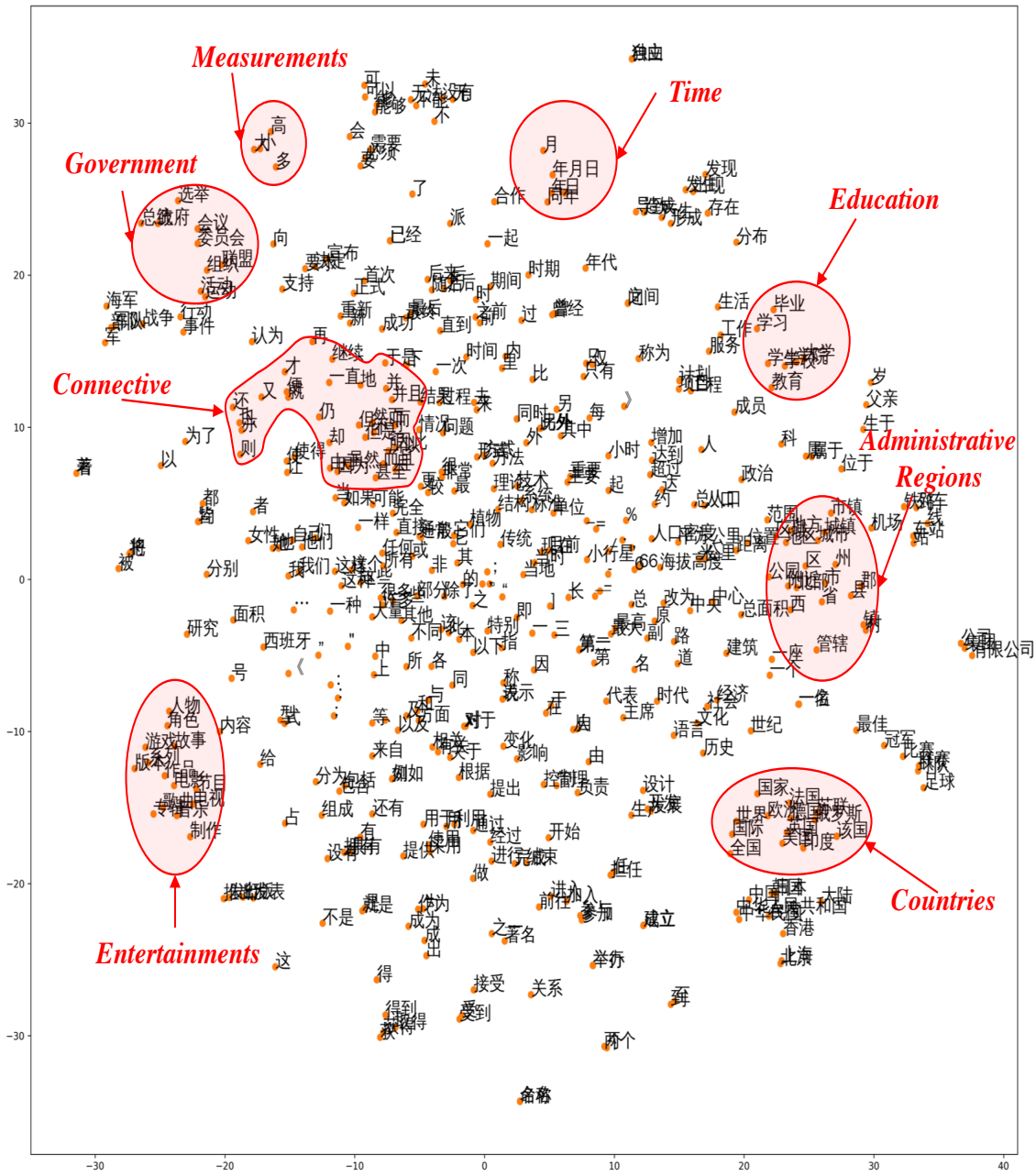


Figure 3.4 Top 500 words in Chinese Wikipedia (till May 2017)

word embeddings trained from Chinese Wikipedia[21] with Gensim (Řehůřek and Sojka, 2010), using the t-SNE dimensionality reduction technique (Maaten and Hinton, 2008) that uses random walks on neighbourhood graphs to allow the implicit structures of all the data to influence the way in which a subset of data is displayed.

In Figure 3.4, I show that several clusters of words of similar semantics or syntactic functions can be identified, as highlighted out in the red shapes. For example, the cluster of countries at the right bottom consists of country names, including France (‘法国’), Russia (‘俄罗斯’), Britain (‘英国’), United States (‘美国’) and India (‘印度’). Right below it is another cluster I did not highlight but is made up of Chinese country name (‘中华人民共和国’, ‘中国’) and big cities such as

[21] https://dumps.wikimedia.org/zhwiki/20171103/zhwiki-20171103-pages-articles-multistream.xml.bz2

Beijing(‘北京’), Shanghai(‘上海’), Hong Kong (‘香港’) and China mainland (‘大陆’). These names of countries and places are mapped close to each other because they are both close in meaning and similar in syntactic roles, i.e. they are all nouns that may be subjects or objects of a topic. In the middle of the plot is a cluster of Chinese connective words, e.g. 还 (‘still’) 仍 (‘still’), 便 (‘then’), 却 (‘but’), 虽然 (‘although’), 由于 (‘since’), 然而 (‘yet’), 但是 (‘but’), 并且 (‘moreover’), 甚至 (‘and even’), 因此 (‘therefore’). They are clustered together because they play more or less the same syntactic role in a sentence. Among these 500 words, I also identified several other clusters of words about education, government bodies, measurement words, entertainments, time units and administrative regions.

Monolingual neural language models represent word meanings via word embeddings built from predictions of word neighbours using the distributed Skip-gram and Continuous Bag-of-Words (CBOW) models proposed by Mikolov et al. (2013a). These models learn word representations that are predictions of the neighbours of a word. Cross-lingual embedding enables the projection of examples available in one language into a shared cross-lingual embedding space and the capability of performing predictions in other languages. To this end, cross-lingual embedding models use a translation matrix (TM) trained on a seed bilingual dictionary to convert monolingual word embeddings into the shared space (Mikolov et al., 2013b). Similar studies aim to improve the process of TM production, e.g. via Canonical Correspondence Analysis (Faruqui and Dyer, 2014), Global Correction (Dinu and Baroni, 2014), or TM orthogonalisation (Artetxe et al., 2016) (Sharoff, 2018). Ruder (2017) gave a nice overview of cross-lingual models and categorised them into four different approaches:

- **Monolingual mapping.** Monolingual word embeddings are trained on large monolingual corpora first. Then a linear mapping in different languages is applied to enable unknown words from SL are mapped to TL.
- **Pseudo-cross-lingual.** By mixing contexts of different languages, a pseudo-cross-lingual corpus is created to train word embeddings.
- **Cross-lingual training.** Word embeddings are trained on a parallel corpus, and the cross-lingual constraint between embeddings of two languages are optimised so that embeddings of similar words are close to each other in the shared vector space.
- **Joint optimisation.** Models are trained on parallel or monolingual data, and a combination of monolingual and cross-lingual losses are jointly optimised.

Unsupervised distributed representations of words capture important syntactic and semantic information about languages, and this technique has been successful in improving MT performance (Zou et al., 2013; Wu et al., 2014; Gouws et al., 2015). Results in these studies show that bilingual translation scores generated

from bilingual embedding models when used as features in SMT systems can achieve significant improvement in terms of BLEU scores. Recently, there are also work in which monolingual or bilingual embeddings are used for MTQE (Shah et al., 2015; Abdelsalam et al., 2016; Paetzold and Specia, 2016; Scarton et al., 2016). Word embeddings are used to represent the texts (sentences or documents) either for direct training with an algorithm (Paetzold and Specia, 2016; Scarton et al., 2016) or computing the similarity between STs and TTs that is consequently used as a feature (Shah et al., 2015; Abdelsalam et al., 2016; Scarton et al., 2016).

I take the second approach. As I am estimating quality for human translations at both sentence and document level, I train word representations directly through a dictionary-based bilingual word embedding mapping (Artetxe et al., 2017). In this approach, a small dictionary of the bilingual lexicon is used as the input of a self-learning system that learns a better mapping and dictionary iteratively until a certain convergence criterion is met. With sentence representations, I compute cosine similarity between STs and TTs with the averaged vectors of the weighted summation of the vectors of all non Out-of-Vocabulary (OOV) words in them.

3.2.2.7 Bilingual Terminology

Two situations often arise in the process of translation and interpreting, i.e. 'the terminology requirement of any translation (terminology in translation)' and 'the translator's (or interpreter's) terminology needs (terminology for translation)' (Cabré, 2010, pp.358). In this sense, terminology plays an important role in the process of translation and interpreting. This is because it helps translators organise their domain knowledge and provides the means (usually terms in various lexical units) to express subject knowledge adequately. In the meantime, lots of bilingual terms come from translations. Translation scholars and practitioners have realised the reciprocal relationship between terminology and translation, maintaining that terminology correctness is associated with the quality of the translation (and interpretation) (Brunette, 2000; Xu and Sharoff, 2014; Kim et al., 2015; Karoubi, 2016).

The acknowledgement of the contribution of terminology to translation quality is also echoed by the translation industry. The 'up-to-date knowledge of the subject material and its terminology in both languages' is written in the ATA Code of Professional Conduct and Business Practices (2009) (ATA, 2010). Accurately reproducing the content of the original and using appropriate terminology has become the official assessment criteria of some in-use translation-error-based evaluation schemes. For instance, the MeLLANGE project (Secară, 2005) defines six terminology errors[22], and the Multidimensional Quality Metrics lists terminology as one of the eight major dimensions, which is subdivided into three children issue types (term inconsistency,

[22]The main terminological errors are incorrect terminology, false cognate, term translated by non-term, inconsistent with glossary, inconsistent within TT, inappropriate collocation, and user-defined errors.

ZH	Glossary
心电 呼吸 系统	ECG Respiratory system
心脏 呼吸 系统	Heart Respiratory system
心脏病 和 血液病	Heart disease and blood disease
心 血管 疾病	Cardiovascular diseases
心脏病 和 循环 系统 疾病	Heart disease and circulatory system diseases

Table 3.7 Examples of Incorrect Terminology Error

termbase[23], and terminology domain[24]) (Lommel et al., 2014a). SAE J2450 also includes the wrong term as one of the default errors (SAE, 2001). These error types and practice code demonstrate that adherence to the specified terminology is considered a central concern in translation for the delivery of quality-assured translations. In my data, I have identified that incorrect terminology has been a common problem for translations unfavourably scored by human annotators. Table 3.7 lists some samples of infelicitous Chinese (ZH) translations for the English term 'cardio-respiratory system' (心肺 系统) highlighted by the annotators.

From a user's expectation perspective, appropriate terminological use is also viewed as one of the important quality parameters. For the purpose of marketing, companies will localise the manuals that accompany their products. Even though they strive to release all language versions of their products at the same time, localisation cannot be done at the expense of quality to endanger the customer satisfaction. Their dissatisfaction will lead to more potential losses in revenue. Therefore, speed and quality are what localisation services users are looking for (Warburton, 2013). They would expect that all the terms are translated correctly and consistently, and translators will not invent terms randomly wherever SL terms cannot find an equivalent in the TL without scientific analysis and sufficient documentation.

It is clear that finding an equivalent for terms in a translation impacts the overall quality of the translation. When assessing a translation, evaluators should also consider how well a translator achieves in successfully rendering those terms in an ST. However, this element of translation has not drawn enough attention from researchers in MTQE, and in human translation quality assessment, the complete evaluation of the translation of terminology is carried out by human assessors and evaluators manually and subjectively, with or without references. Manual compilation of bilingual term lists for each translation evaluation task is an expensive and laborious effort, hence the rarity of an up-to-date, specialised and relatively comprehensive term database for the TQE purpose.

I propose to automate as much as possible the process of recognition and evaluation of terminological equivalence in STs and TTs. Thus, features related to bilingual terms as quality indicators are explored. I propose a supervised learning method with minimal requirement for linguistic processing (Details will be provided

[23]a term is translated with a term nonconforming to the specification.
[24]a term is translated as a term from a different domain.

ID	Definition	Feature
F181△	log ratio of shallow features	7
F184△	shallow features CBD	3
F185-204△	log ratio of ST-TT cohesion and coherence features	20
F205-206△	ST-TT cohesion and coherence features CBD	2
F207-243△	log ratio of dependency features	37
F244 △	dependency feature CBD	1
F245-250 △	log ratio of constituency features	6
F251 △	ST-TT constituency CBD	1
F252-263△	log ratio of PoS tags	12
F264 △	PoS tag CBD	1
F268△	log ratio of semantic labels	4
F269△	Semantic role labels CBD	1
F270-275△	alignment features	2*3
F276-312△	MT Back-translation features	37
F313-349△	Pseudo-reference features	37
F350-351 △	terminology precision and recall	2
F354 △	log ratio of ST-TT language model features	3

Table 3.8 Bilingual Features

in Chapter 4) to automatically recognise terms from the bilingual running texts. Two features will be combined with other features for HTQE.

- normalised ratio of automatically identified terms to the length of TTs
- normalised ratio of automatically identified terms to the length of STs

Bilingual features are listed in Table 3.8.

3.2.3 Language Modelling Features

As far as language modelling features are concerned, they consist of TT lexical language model (LM) perplexity score and TT POS LM perplexity score. These features will be explained in the following.

3.2.3.1 LM Perplexity Score

Language models are statistical models of word sequence. They play an essential role in many NLP applications, such as MT, speech recognition, handwriting recognition, language tagging and parsing (Jurafsky and Martin, 2008, pp.4). A statistical language model is a probabilistic distribution over sequences of words. Given a model of length i, it is in a sense exhaustive of vocabulary of the target language, and it can assign to a sequence of words a joint probability

$$P(W) = P(w_1, w_2, \ldots, w_i),$$

where $P(W)$ can be computed using the chain rule as

$$\begin{aligned} P(w_1\, w_2\, \ldots w_n) &= P(w_1)\, P(w_2|w_1)\, P(w_3|w_1, w_2) \ldots P(w_n|w_1, w_2 \ldots w_{n-1}) \\ &= \prod_{i=1}^{n} P(w_i|w_1\, w_2\, \ldots w_{i-1}). \end{aligned}$$

We can then estimate the joint probability of an entire sequence of words by multiplying together a number of conditional probabilities. Given an n-gram model, we can approximate its history by just the last few words. With a bigram model, we can rewrite $P(W)$ as

$$P(W) = P(w_n|w_{n-1}) \approx \prod_{i=1}^{n} P(w_i|w_{i-1}).$$

In a similar vein, Given a trigram model, $P(W)$ is calculated in the form of

$$P(W) = P(w_n|w_{n-1}, w_{n-2}) \approx \prod_{i=1}^{n} P(w_i|w_{i-2}, w_{i-1}).$$

To compute the specific n-gram probabilities, MLE is employed to get the MLE estimate paramters of a n-gram model. For word x and y in any corpus C with words w, the probability of bigram xy can be computed and normalised by the sum of the all the bigrams sharing the first word x:

$$P(y|x) = \frac{C(xy)}{\sum_w C(xw)},$$

which is then further simplified as

$$P(y|x) = \frac{C(xy)}{C(x)}.$$

In building a language model, data sparsity is a major obstacle as word sequences may not occur in the training data. Backoff-smoothing (Katz, 1987) is often adopted to tackle the issue. In this study, I use the shift-beta smoothing on IRSTLM (Federico et al., 2008). LM is a technique that estimates the conditional probability of a word given its history in the n-gram. Thus, an n-gram model is the common strategy to model the probability distribution of sequences. It is assumed that the identity of the $i'th$ word in the sequence depends only on the identity of the previous n-1 words (Collins, 2013).

LM perplexity score is a canonical measure of goodness of a statistical language model (Azzopardi et al., 2003). It is defined as the inverse probability of the distribution W, normalised by the number of words:

$$PP(W) = P(w_1 w_2 \cdots w_N)^{\frac{1}{N}},$$

where $\mathrm{P}(\cdot)$ is LM probability of the test set (Jurafsky and Martin, 2008, pp.14). Thus if we are computing the perplexity of W with a trigram language model, we have

$$\mathrm{PP}(W) = \sqrt[\mathrm{n}]{\prod_{\mathrm{i}=1}^{\mathrm{N}} \frac{1}{\mathrm{P}(w_{\mathrm{i}}|w_{\mathrm{i}-2}, w_{\mathrm{i}-1})}} \quad .$$

The language modelling and ML community has been using low perplexity representations to indicate the likelihood of a sequence of words. It is very useful to model a prior distribution over sequences of words and tell which are probable or unlikely in a language. In order to automatically rank the translation produced in terms of fluency, a statistical language model of the target language is often built and then applied to judge the probability and perplexity of the target text. The higher ranking output with low perplexity score is deemed to be the more fluent and therefore a better translation.

3.2.3.2 Log Probabilities

In most cases, language model probabilities are represented and computed in log format in that it can efficiently handle the problem of numerical underflow. Since probabilities are less than or equal to 1, the more probabilities we multiply together, e.g. in the case of a very long sentence, the smaller the final product becomes. Using log probability allows us to get numbers that are not too small and do computation and storage in an efficient manner (Jurafsky and Martin, 2009). I compute the log probabilities for each sentence and each document, using the 3-gram language model trained with the selected English and Chinese Monolingual corpus (described in Chapter 5).

3.2.3.3 Out-of-Vocabulary Words

Unknown words, or out of vocabulary (OOV) words, are words that we do not see in a language model. Given a large size of training data, if words in a translation are not commonly seen in the language model, it is very likely that either these words are too domain-specific or they are simply wrongly used or made up. Thus, high percentage or normalised counts of OOV words to some extent indicate a deterioration of translation quality, or viewed positively they indicate the lexical variation and creativity of translators.

N-grams models have received intensive research since its invention. Sequence n-gram model is one of the typical extensions to traditional n-gram model. A sequence can be a sequence of words, word classes, POS tags or whatever symbols bear lexicogrammatical information. In this study, a translation lexical LM probability, i.e. the probability of the sequence of words themselves and the probability of sequences of POS tags are measured against the pre-trained language models of word n-grams and POS n-grams. I set the n to be 3 as it is a common choice with

ID	Definition	# Feature
F356	ST & TT 3-gram LM perplexity score	2
F358	ST & TT number of Out-of-Vocabulary (OOVs)	2
F 360	ST & TT PoS 3-gram LM log probability score	2

Table 3.9 Language Modelling Features

large training corpora (millions of words) in practical NLP applications (Rosenfeld, 2000).

Language modelling features selected are listed below in Table 3.9

3.3 Summary

The purpose of the current chapter is to design a framework of the feature set for the representation of human translation data for later ML tasks.

Following the survey of features used for MTQE in Chapter 2, the author proposes to reduce the constraints on translations to linguistic ones only and treat translation as a linguistic transfer activity in which all external constraints are embodied in the final products themselves. Thus, selecting linguistic and shallow features that are comparatively easy to implement and including them in the framework is a viable solution.

The proposed feature set consists of monolingual, bilingual and language modelling features, with heavily extended new features targeting specifically HTQE to overcome the observed fallacies of current MTQE. Except for those shallow features, I have adapted and designed an extended collection of features (marked with triangle in Tables 3.5, 3.8 and 3.9) for HTQE. In the next chapter, I present the solution to bilingual terminology extraction for HTQE.

Chapter 4

Cross-lingual Terminology Extraction for HTQE

4.1 Introduction

Automated Terminology Recognition (ATR) aims to identify terminological units in domain-specific corpora. Such information is useful for several tasks, such as dictionary compilation, ontology building, information retrieval, text summarisation (Conrado et al., 2013). Research in ATR for the TQE purpose is still at an earlier stage (Oliver, 2017), partially because it deals with bilingual terminology extraction, which is believed to be more complicated than monolingual term extraction (Gaussier, 2001) .

For efficient term extraction, many specific methods have been proposed. Kang et al. (2005) have identified that **linguistic**, **statistical** and **hybrid** systems are three predominant approaches to ATR. Linguistic systems make use of POS tags, lexicons, syntax or other domain- and language-specific linguistic structures. However, the rule-based linguistic approach (Ananiadou, 1994; Bourigault et al., 1996; Heid, 1998; Wermter and Hahn, 2005; Fahmi et al., 2007) is heavily language-dependent with low portability and extensibility to a different language. Another critical drawback of linguistic systems is their incapability of identifying boundaries of complex and nested phrases, i.e. the Nested Noun Phrases challenge (Li et al., 2012b). Most rule-based systems execute pattern clauses in order. The process may introduce a chunk boundary that prevents a latter pattern from executing. Take the phrase 'Provisions of shading devices' for example. In the case of ontology learning, the nested phrase 'shading device' is preferred over the longer phrase. However, the dilemma occurs when it comes to the noun phrase 'the authorities having jurisdiction'.

Purely statistical systems are commonly achieved by utilizing frequency, significance and degree of association, and heuristics measures to determine the termhood of words and the unithood. These methods concentrate on real terms at the top N ranks of total candidates in an attempt to facilitate subsequent human experts validation and filtering (Kageura and Umino, 1996; Zhang et al., 2008). The

major advantage of the statistic approach is language-independence. However, studies have shown that statistics, such as frequency, Mutual Information and its variants, have incongruous performance on different datasets and behave significantly different to favour either high-frequency events (i.e. 'noisy') or the rare events causing a high number of terms identified (i.e. 'silence'). Quantity and quality of the dataset also have been identified as the important factors influencing statistical approaches (Li et al., 2012a).

The predominant hybrid approach exploits the advantages of both rule-based and statistical methods. Statistical steps are applied to the narrowed-down list of candidate terms identified by various domain-specific linguistic heuristics so as to further improve the accuracy. Nevertheless, by nature, the combination of linguistic filters and statistical ranking would lead to degenerated precision with the increase of recall, as reported in (Pazienza et al., 2005).

Terminology extraction in translation involves cross-domain and cross-language datasets. However, compiling linguistic rules to extract terms from both languages is laborious and costly, and the datasets in different languages often cause the incongruous performance of statistics. Therefore, it would not be viable to take the hybrid approach for an economy purpose, i.e. manual preparation of linguistic filters to be used with statistical extractors is costly. Thus, I follow the machine learning approach and present a supervised learning approach for term extraction based on language-independent features. I present a quality-oriented approach focusing on terminology for translational data. The whole pipeline is illustrated in Figure 4.1. A range of representative and language-independent algorithms are exploited to compute term representations to train classifiers that classify n-grams into terms and non-terms. The approach I take differs from other ML approaches based on linguistic features and context information (Erdmann et al., 2009; Zhang et al., 2010; Li et al., 2012a; Hakami and Bollegala, 2017). Instead, I adopt a feature set based on language-independent statistical measures that can be computed directly from the corpus containing target terms. Only minimal linguistic processing such as tokenisation and lemmatisation is used for data and feature extraction. Another contribution of this research is that I carried out experiments to provide comparative results, based on a standard benchmark of publicly available corpora across four domains and two languages (Chinese and English). Besides, unlike most of the previous research that focuses on Indo-European languages, such as English, Spanish, Swedish and Russian, which are closely related, I work on a distant language pair, namely English and Chinese.

4.2 Related Work in Bilingual Term Extraction

Taking a 'similarity context vector approach' (Déjean and Gaussier, 2002), which associates the words with the context vectors of the nearest lexical units in the bilingual

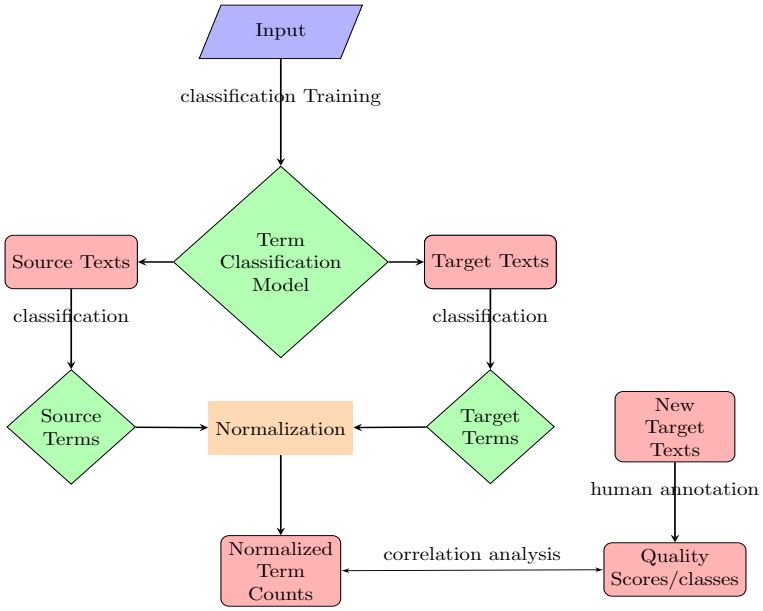


Figure 4.1 Terminology-focused Translation Quality Evaluation Pipeline

lexicon, Daille and Morin (2005) propose tackling fertility[1], non-compositionality[2] and variation[3]. They use both linguistic and statistical methods to identify the source and target terms from monolingual corpora and then link multi-word terms (MWTs) in the source language to single words and multi-word terms (MWTs) in the target language. The authors claim their approach has the capacity of handling French-English terms variations. The method has demonstrated satisfactory performance of identifying translations (within the top 20 candidates) for multi-word terms. However, lower accuracy (i.e. candidate terms are far out) in identifying one-to-one (i.e. single word term for single word term translation) and one-to-many (i.e. single word term and multi-word term for single word term translation) is observable. It is also questionable to what extent this variation-oriented approach applies to other languages which are less morphosyntacticly variational, such as Chinese and Japanese.

Based on the same idea of context similarity (Rapp, 1995; Fung and Yee, 1998), Saralegi et al. (2008) use Earth Movers Distance (EMD). In the case of document similarity, the EMD is employed to compute the minimum cost of the weighted graph as the similarity value between two documents. Documents are represented as the word distributions and weighted on their individual $\mathrm{TF} * \mathrm{IDF}$ values. On such a basis, a weighted graph is constructed to find the flow between words in document A and B with the minimal cost. The resulted work normalised by the total flow denotes the EMD (Wan and Peng, 2005) to compute the similarity between corpora. A bilingual dictionary is used to translate the source context vectors for computing the context vector similarity. In the end, the authors find combining cognates in the ranking

[1]Terms are not always translated by terms of the same length.

[2]When terms are translated, target sequences of terms are not typically composed of translations of their parts.

[3]Terms could appear in texts under different forms reflecting either syntactic, morphological or semantic variations (Daille, 2003).

process more suitable for texts of science domain where cognates are prevalent. The authors also admit that the relatively small sizes of their test corpora cause a significant decrease in the recall, as very few words pass the frequency threshold necessary to obtain a good precision in the phase of computing context vector similarity.

Erdmann et al. (2009) propose the extraction of bilingual terminology from large multilingual Wikipedia to complement bilingual dictionaries for languages and text domains where no parallel corpora exist. They develop a method that analyses redirect page and anchor text information to extend the number of term-translation pairs while maintaining high accuracy, by training a Support Vector Machine (SVM) classifier on a portion of randomly extracted and manually annotated term-translation pairs. The trained SVM classifier further increases the recall of finding translation term pairs. However, the precision still warrants improvement in comparison to other methods (e.g. online dictionary BEOLINGUS[4] method, Interlanguage link method).

Lu and Tsou (2009) deal with bilingual term extraction in a comparable corpus of patents. The authors present a framework consisting of 4 main steps: monolingual extraction of single-world terms (SWTs) and multi-word terms (MWTs), parallel sentence identification, bilingual candidate term extraction and bilingual term classification. The proposed framework provides an alternative to mining the bilingual terms from sentence-level parallel corpora, but the extraction of monolingual candidate terms is limited to noun phrases (< 5 words) from comparable monolingual corpora. Thus, other terms, e.g. verb and adjective terms, are often ignored.

Vintar (2010) describes a hybrid term extractor using morphosyntactic patterns and statistical ranking to select domain-specific expressions for each language pair. Term translations between two languages are identified using the bag-of-words approach. The proposed method does not require parallel corpora and hence applies to comparable monolingual corpora, provided a domain-specific lexicon or a bilingual lexicon is ready. Despite all the advantages of handling term variations and multiple translations, high precision and its applicability to automatic term recognition (ATR) from non-parallel corpora, the system still depends on the manual compilation of monolingual pattern rules. As a consequence, it suffers from recall problem[5] of candidate terms from monolingual domain-specific corpora.

TTC TermSuite (Rocheteau and Daille, 2011) is a 4-step graphical utility based on UIMA framework (Ferrucci and Lally, 2004) to perform bilingual term extraction from comparable corpora. The tool consists of 4 procedural steps (i.e. text preprocessing, linguistic analysis, monolingual term extraction and bilingual term alignment) and supports analysing large volumes of unstructured data. The tool provides a processing chain for cross-language term extraction from comparable corpora, but it requires domain experts to compile pattern rules during the linguistic

[4] http://dict.tu-chemnitz.de/

[5] Because manually compiled syntactic patterns usually do not have complete coverage of candidate terms and often introduces noise as well.

analysis phase to configure both the source language and target language before term extraction and alignment. Even worse is that the precision of extraction and alignment generally decreases with the increase of recall, as is often the case with hybrid term extraction method (Zhang et al., 2016).

Gaizauskas et al. (2015) developed a multi-component (data collection, domain classification, monolingual text extraction and bilingual term alignment) system that automatically gather domain-specific bilingual term pairs from web data. Their workflows consist of four steps in sequence: collecting parallel and/or comparable corpora; classifying the collected web resources into pre-defined categories of domains via vector representation; extracting terms from monolingual documents in different domains through linguistic and frequency filtering; aligning the terms in parallel or comparable document pairs using context-independent mapping. Four human assessment tasks and associated protocols are also designed to evaluate the term extraction pipeline. Results show that 94% of the aligned terms were correct translation equivalents, and there is not much variance between the accuracies of terms with varying lengths. Despite the favourable findings from the evaluation process, one of the most significant concerns is that the method mainly works with data in European languages, and no clear evidence the pipelines will work on remote language pairs.

In order to build a bilingual medical lexicon for the multi-lingual and multi-national treatment, Xu et al. (2015) attempt to extract a bilingual lexicon from English and Chinese discharge summaries with a small seed lexicon. Label propagation (LP) method is adopted to extract SWT translation pairs, and MWTs are dealt with using term alignment [6]. In comparison to a baseline based on simple context similarity, the proposed method shows superior performance with improved accuracy. However, the translation generation by the combination method presupposes that the compositionality is ubiquitous among the bilingual term pairs. This phenomenon, however, does not always hold true (Daille and Morin, 2005).

Hakami and Bollegala (2017) train a binary classifier to determine whether two biomedical terms written in two languages are translations. Several feature space concatenation methods (e.g. linear concatenation, pair-wise concatenation) are proposed to overcome the lack of common features between language pairs. Several aspects of the performance of the proposed methods can be further improved. Specifically, the ambiguity of term translations and identification of synonymous terms need to be further addressed.

Though the aforementioned studies have attempted to deal with the task of identifying bilingual terms from different resources, these systems and pipelines do not readily serve the purpose of finding term pairs from the translated texts to be evaluated. Besides, these term extraction methods often involve the compilation of

[6] Translation candidates of each component word of a Chinese MWT are first obtained and then used to generate combinations for the whole MWT, upon which some plausible candidates are selected.

linguistic rules and rely on external tools or resources, and some of the methods have a narrower scope as a result of focusing on specific types of terms (e.g. MWT or NPs). As our aim to evaluate how well terms are translated[7] in students' translations on different topics from various domains, we can neither manually design generic linguistic patterns for all translations nor tolerate low recall for the sake of precision. Therefore, a method of automatically identifying terms from both STs and TTs to quantify their correspondence relation is needed.

For this purpose, I come up with an ML technique that uses language-independent features to train a classifier which classifies n-grams into terms and non-terms in both STs and TTs and finds their correspondence relation. In the following, I describe the proposed method.

4.3 Terminology Classification

The following is a brief description of the features I use to train the term classifiers.

4.3.1 Common Statistics as Features

This study is based on the assumption that domain-specific terms have morphological feature, distribution feature, context feature, domain-specific feature and so forth, which distinguish them from common words. I assume that a selection of such representative features can help build a feature space that supports further classification. The motivation of feature selection is to find one or more features that can provide an efficient representation of the candidates from a given corpus, with a particular focus on independence from languages and domains. Moreover, I hope that the representations are effective to facilitate extracting both SWTs and MWTs.

From a pragmatic point of view, the features used in this research are computed by JATE 2.0 (Zhang et al., 2016), a recently released open-source ATR tool. JATE 2.0 implements 10 representative statistical ATR algorithms that are scalable to large corpora, including global distribution weighting schemes, e.g. TF-IDF, heuristic based domain-specificity measures, e.g. C-Value (Frantzi and Ananiadou, 1996), significance of association, e.g. χ^2 (Matsuo and Ishizuka, 2004), contrastive measures for domain-specificity, e.g. Weirdness (Ahmad et al., 1999), graph based measures, e.g. RAKE (Rose et al., 2010) and ensemble based measures unifying both unithood and termhood measures, e.g. GlossEx (Park et al., 2002). These features and their corresponding algorithms are listed in Table 4.1.

TTF, namely Term Total Frequency, is the total frequency of a candidate in the target corpus. This algorithm is first documented in Justeson and Katz (1995), where frequency information is taken into account for retrieving 'words or phrases

[7] To see if all terms are translated correctly, or if good translations have more terms translated and poor translations have fewer.

Feature	Algorithm	Reference
TTF	Total Term Freqeuncy	(Justeson and Katz, 1995)
ATTF	Average Total Term frequency	–
TTF-IDF	TTF with Inverse doccument Freq.	–
RIDF	Residual IDF	(Church and Gale, 1999)
C-Value	C-Value	(Frantzi and Ananiadou, 1996)
RAKE	Rapid Keyword Extraction	(Rose et al., 2010)
χ^2	Chi-square	(Matsuo and Ishizuka, 2004)
Weirdness	Weirdness	(Ahmad et al., 1999)
GlossEx	Glossary Extraction	(Park et al., 2002)
TermEx	Term Extraction	(Sclano and Velardi, 2007)

Table 4.1 Features Used for Term Extraction

that are both highly indicative of document content and highly distinctive within a text collection'.

Thus, in a relatively small size document, n-grams occur two or more times are likely to be term units. Accordingly, the frequency of candidate terms in the running text places an important constraint on the terms it contains.

ATTF takes the average of TTF and divides it by the number of documents in which the candidate term occurs.

TTF-IDF is adapted from the classical TF-IDF. It replaces the local distribution measure with the global distribution across the whole corpus and attempts to assign higher values to words that appear more frequently in a few documents across the whole corpus. This measure generally works well to filter common words (Ramos et al., 2003).

However, like most of the frequency-based measures, it also suffers from some inherent conventional drawbacks (Xia and Chai, 2011). Specifically, terms appearing intensively on a few documents may not represent the specificity of terms in the domain, when considering a corpus containing several subsets of documents about different subtopics. Moreover, incorrect weighting is more likely to yield many uninformative words such as function words, auxiliary words and conjunctions. This side effect can be addressed by removing stopwords. Thus, TTF-IDF often serves as a baseline of termhood metrics or is adopted as one of the several features in ATR.

RIDF, known as residual IDF, was first proposed to identify keywords in a collection of documents and then adapted in ATR as a termhood metric. This measure captures the deviation of the actual IDF score of a candidate from its expected IDF score on a Poisson distribution, of which real term (or keywords) is assumed to be higher than non-term (or ordinary words).

C-value is a typical hybrid ATR approach that combines linguistic and statistical information, and more importantly, it can provide an ensemble representation of unithood and termhood. Such a measure considers the impact of frequency and length of a candidate term, capable of enhancing the conventional statistics of

frequency. Thus, it is sensitive to nested terms such as the candidate term 'T cell' nested in longer terms 'peripheral blood T cell', 'naive T cell' and 'T cell activation'.

Although it is initially proposed to extract MWTs, C-value demonstrates the flexibility to handle shorter MWTs and even SWTs

RAKE, short for Rapid Automatic Keyword Extraction, is a type of graph-based ranking model based on word co-occurrences. RAKE uses stop words/phrases and punctuations/delimiters to isolate keywords from a document. The co-occurrence of words in MWTs is leveraged as a meaningful clue to determine the importance of a candidate term. This metric avoids the common drawback of the purely frequency based statistics, i.e. bias towards SWTs. RAKE metrics can evaluate the exclusivity, essentiality and generality of extracted candidate terms. The measurement is based on three metrics, including word frequency, word degree (the occurrence of a word in more extended candidate MWTs) and the ratio of word degree to frequency.

χ^2 measure is commonly used for bigram statistics. The observed frequency of co-occurrence is represented in a matrix, by which a null hypothesis can be tested whether bigram tokens co-occur by chance. JATE 2.0 adapts the measure to work with both SWTs and MWTs. If a term has no co-occurrence information, a zero score is assigned.

Accordingly, a larger χ^2 means that the word is more important in the document. χ^2 is known to be biased towards low-frequency words. Other factors such as frequency are usually combined to counteract its deficiency.

Weirdness, or **specificity**, is a type of contrastive ranking technique, which is particularly interesting for identifying low-frequency terms. The technique was introduced by Ahmad et al. (1999) and based on the hypothesis that the distribution of a term within one domain differs from the distribution of the same word (s) in other domains or general language use. It often compares the frequencies of candidates in a domain-specific corpus with those in a reference corpus.

For MWTs, a geometric average of the weirdness of each component word is computed (Knoth et al., 2009). British National Corpus (BNC)[8] is adopted as the general purpose reference corpus in JATE 2.0. The corpus is used in other two contrastive ranking related algorithms, i.e. GlossEx and TermEx.

GlossEx was put forward to rank all glossary items of varying lengths. This is another hybrid approach which measures the goodness of a term by combining term specificity (i.e. termhood) and term association (i.e. unithood). The former quantifies how much an item is related to a specific domain, and the latter describes the association degree of words in the term.

The algorithm is a combination of two components with different weights, including term-specificity (TD) and term cohesion (TC). Essentially, the TD is a contrastive approach that computes termhood by comparing the relative probability of a term against a reference corpus. TC is a generalised Dice Coefficient that measures the degree of co-occurrence of words in MWTs. Like all other co-occurrence measures,

[8] http://www.natcorp.ox.ac.uk

this metric is biased towards SWTs. Thus, the weight to TC (i.e. β value) is set to 0.1 for SWTs in JATE 2.0 to address the issue.

TermEx is very similar to GlossEx with extra extension of entropy-related Domain Consensus (DC) metric. DC gives more weight to a term that has even probability distribution across the documents of the domain corpus. Another two components are the Domain Pertinence (DP) and Lexical Cohesion (LC), which are essentially the same as Weirdness and TC in GlossEx respectively. The final algorithm is a linear combination of the three metrics with adjustable weights (default to 1/3 in JATE 2.0).

4.3.2 Training Monolingual Term Classifier

4.3.2.1 Corpus

6 corpora are selected in my experiment, covering four different domains and two different languages (of varying sizes). The GENIA corpus (Kim et al., 2003) is a collection of biomedical documents, and it is the most popular dataset used in ATR. ACL RD-TEC (Version 1.0) (Q. Zadeh and Handschuh, 2014) is a dataset created recently for evaluating the extraction and classification of terms from literature in Computational Linguistics. The first version is used in my experiment, containing over 10,900 scientific publications and 82,000 manually annotated terms.

TTC, short for Terminology Extraction, Translation Tools and Comparable Corpora, a recent European project covering eight languages, contributes various linguistic resources for bilingual term acquisition and translation (Blancafort et al., 2010). Two English-Chinese (EN-ZH) comparable corpora, totalling four datasets for two specialised domains in Wind Energy (TTC-W) and Mobile technology (TTC-M), are used in my experiment as test sets.

I present the detailed information of all 6 corpora in Table 4.2.

Corpus	# of documents	Size(tokens)	Reference Term List
GENIA	1,999	420,000	35,800
ACL RD-TEC	10,900	36,729,513	22,013
TTC-W (EN)	172	750,855	188
TTC-M (EN)	37	308,263	143
TTC-W (ZH)	178	4,263,336	204
TTC-M (ZH)	92	2,435,232	150

Table 4.2 Corpora Used for Training Term Classifier

4.3.2.2 Dataset Pre-processing

Four English corpora are first tokenised by JATE 2.0 OpenNLP English tokeniser and sentence splitter. Two Chinese TTC corpora are pre-segmented by white space. Therefore, the Solr white space tokeniser that comes with JATE is directly applied.

Next, I focus on the n-gram candidate terms with a maximum length of 5 ($1 \leqslant \mathfrak{n} \leqslant 5$) in the experiment. I then further filter these n-gram candidates by removing stop words[9] and setting minimum term frequency ($\geqslant 2$) and character length range for English ($2 \leqslant \mathfrak{l} \leqslant 50$). ASCII folding and English lemmatisation are also performed for English corpora. Ten aforementioned ATR algorithms are then run separately to score all the candidates, the output of which will be used as features to the subsequent process of dataset generation.

In the final step, training data and testing data are processed by n-gram string matching with ten features outputted separately by the ten algorithms. The n-gram datasets are further matched with specific Reference Term List (RTL) from each dataset. Any matched n-gram will be labelled as true positive and those having no matches will be viewed as non-terms. By this way, I eventually have 4,240 true terms from GENIA and 9,057 from ACL RD-TEC respectively. To examine the effect of different sizes of training and testing data, GENIA and ACL RD-TEC datasets are used as training data in the experiment, and they are also used to validate classifiers trained on each other. See Table 4.3 and Table 4.4 for the details of the n-gram datasets generated in the experiment.

N-gram Datasets	# of terms	# of non-terms
GENIA	4,240	45,350
ACL RD-TEC	9,057	858,544
TTC-W (EN)	120	30,925
TTC-M (EN)	149	20,505
TTC-W (ZH)	125	132,407
TTC-M (ZH)	168	105,599

Table 4.3 Terms and Non-terms in Ngram Datasets

N-gram Dataset	Size	Training	Testing
GENIA	49,590	Y	Y
ACL RD-TEC	867,601	Y	Y
TTC-W (EN)	31,045	N	Y
TTC-M (EN)	20,654	N	Y
TTC-W (ZH)	132,532	N	Y
TTC-M (ZH)	105,764	N	Y

Table 4.4 N-gram datasets generated in experiment

It is also commonly seen in the field of ATR that terminology datasets are highly imbalanced, e.g. fewer positive and much more negative instances. Negative n-gram candidates, i.e. non-terms, have a highly unbalanced distribution in the dataset as shown in Table 4.3. To mitigate this effect in the subsequent training stage, I apply the undersampling method (Lemaître et al., 2017) to the majority of non-terms for the training set.

[9] English n-grams are filtered by SMART stopword list, available via https://goo.gl/DXwrgy; Chinese n-grams are pre-filtered by both English stopwords and the 125 Chinese stop words compiled by Kevin Bouge, available via https://goo.gl/a0gRzd.

4.3.2.3 Setup

The Random Forest (RF) learning algorithm in Scikit-Learn (alias sklearn) (Pedregosa et al., 2011) is chosen to train a classifier. RF does not expect linear features and can handle very well high dimensional spaces and a large number of training examples (Breiman, 2001). In the training stage, all training sets are split proportionally (75% for training and 25% for testing). For model selection, stratified ten-fold cross-validation is used and repeated grid-search is employed for parameter tuning. Also, each individual feature is scaled to the range $[-1, 1]$ with the MaxAbsScaler API[10].

4.3.2.4 Evaluation Methods

The performance of classifiers trained on two training datasets ('GENIA' and 'ACL RE-TEC') is evaluated on the held-out datasets and other 5 separate testing datasets.

Although the task is treated as a binary classification problem, I only focus on the evaluation results corresponding to 'term' class, using the standard Precision (P), Recall (R) and F-measure (F1) to measure the model performance.

Table 4.5 presents the previous state-of-the-art methods on four English corpora. First, TTC TermSuite v2.2[11] (Rocheteau and Daille, 2011) is used in the experiment as the primary baseline for four English datasets. At the time of writing, it does not support Chinese processing. POS-based C-Value implementation in JATE 2.0 (Zhang et al., 2016) is also chosen as a baseline for ACL RD-TEC and GENIA corpus. Zhou and Su's system (2004) was the best performant system in the shared task of BioNLP/NLPBA 2004 which uses GENIA as a dataset. These studies only report results with their Top N[12] ranked subset performance. On the same testing datasets, I compare classifiers trained with the training datasets of ACL RD-TEC and GENIA respectively.

4.3.2.5 Term Classification Models

The performance of the trained classifiers on 6 datasets is presented in Table 4.6. The classifiers with the best F1 scores are considered as the best models in the experiment. With regards to the overall recall, baseline results of four English corpora overall are relatively lower than the classifiers trained, except that the recall of Zhou and Su (2004) on GENIA is about 25% higher than that of the model trained with ACL RD-TEC dataset.

[10] http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MaxAbsScaler.html#sklearn.preprocessing.MaxAbsScaler

[11] http://termsuite.github.io/

[12] TermSuite and the similar studies in ATR, especially those linguistic and statistical ones, tend to report the performance of their methods by evaluating the top-ranked results. For example, the precision of the first 10, 20, 100 system recommendations, etc.

		Precision							Recall
Baselines	Dataset	Top 50	Top 100	Top 500	Top 1000	Top 2000	Top 10000	Overall	Overall
TermSuite v2.2	ACL RD-TEC	0.12	0.09	0.15	0.11	-	0.06	-	0.15
	GENIA	0.48	0.46	0.43	0.44	-	0.46	-	0.1
	TTC-W(EN)	0.4	0.29	0.12	0.08	-	0.01	-	0.44
	TTC-M(EN)	0.32	0.24	0.12	0.07	-	0.01	-	0.62
JATE 2.0 CValue (PoS)	ACL RD-TEC	0.46	0.41	0.36	0.35	0.36	0.28	-	0.74
	GENIA	0.94	0.91	0.86	0.82	0.77	-	-	0.1
Zhou & Su (2004)	GENIA	-	-	-	-	-	-	0.76	0.69

Table 4.5 Baselines on Four English Corpora

	GENIA			ACL RD-TEC		
Testing Dataset	Precision	Recall	F1	Precision	Recall	F1
GENIA/ACL(held-out)	0.80	0.84	0.82	0.84	0.88	0.86
TTC-W(EN)	0.79	0.7	0.75	0.84	0.51	0.64
TTC-M(EN)	0.77	0.74	0.75	0.83	0.68	0.75
TTC-W(ZH)	0.58	0.69	0.63	0.67	0.53	0.60
TTC-M(ZH)	0.57	0.60	0.58	0.69	0.51	0.59
ACL RD-TEC(1.0)/GENIA	0.51	0.99	0.67	0.82	0.26	0.40

Table 4.6 Model Performance on 6 Testing Datasets

Dataset	Domain	Passages	# of sentence	Length
XENO	Xenotransplantation	50	14	$234 \sim 473$
SUIBE	Patent	42	11	$297 \sim 376$

Table 4.7 Statistics for Two Trainee Translation Datasets

The precisions of the models trained on the ACL RD-TEC training data, when tested on two TTC English and Chinese testing set, are higher than those trained on the GENIA training data by a margin of 4%-12%, while their recalls on four testing datasets are much lower than those of GENIA-based models by a difference of 6%-19%.

As expected, the Top N precisions of statistic-based baselines (TermSuite v2.2 and JATE 2.0 CValue) decrease gradually with the increase of recall. The overall precisions of the models trained with either GENIA or ACL RD-TEC dataset are much higher than all the Top N subset precisions obtained by TermSuite baselines. In addition, the best precision (84%) of ACL RD-TEC based optimal model on GENIA testing set is much higher than all the top N precisions of JATE 2.0 CValue baseline for ACL RD-TEC corpus by a margin of 38%, 43%, 48%, 49%, 48% and 56% respectively. However, the best precision (80%) of GENIA based model on GENIA testing set is lower than all subsets of Top 1500 precisions of JATE 2.0 Cvalue baselines (by 14%, 11%, 6% and 2% respectively), despite the fact that this result is still slightly higher than the previous best system (Zhou and Su, 2004) by 4% and much higher than all Top N precisions of TermSuite baseline.

4.4 Experiment

In order to find out to what extent translating terminology contributes to human translation quality. I implement terminology recognition on two human translation datasets and investigate how the number of terms and their translations correlates with human annotated quality scores.

4.4.1 Translational Data

In the following, I describe the data used. The first dataset consists of 50 trainee translators' translations of a short passage about xenotransplantation (280 words). The second dataset is a course summative work from Shanghai University of International Business and Economics (SUIBE). There are 42 translations for a rotatory closure design patent in the dataset. I choose these two datasets because they are all trainee translations and contain very domain-specific words that are potentially challenging for trainees. Hereinafter, they are referred as the XENO data and SUIBE data. The basic statistics of both datasets are available in Table 4.7.

As the XENO dataset is part of the QE dataset, it has been annotated on four dimensions ranging from content transfer (UT), terminology (TS), idiomatic writing (IW) and target language conventions (TM). More detailed information about the annotation is provided in Chapter 5 of this thesis.

As regards the SUIBE data, I use the original scores assigned by the course instructor who also took into consideration students' in-class performance and the formality elements of the translation task. This suggests that the scores are not entirely based on the translations themselves. All the scores are on a percentile scale, and no further processing was done.

4.4.2 Term Count Normalisation

The normalisation process aims to relate the term counts in the TTs to the terms in the STs so that TTs can be measured and compared across different translations. I assume that a higher relative number of terms counts indicates a more successful translation regarding term adequacy, which in turn contributes to the overall translation quality text-wide. The purpose of this normalisation process thus is to obtain a form of term count that is comparable within translations of different lengths from STs containing a different number of source terms. In the following experiment, I compute the normalised term count for each translation at the document level. The normalised term count in TTs is calculated using Eq. 3.1.

4.4.3 Evaluation

To investigate whether the automatically identified terminology is related to the quality of the trainees' translations, I compute and report the Pearson correlation coefficient (r) (the primary criterion), Spearman rank correlation coefficient (ρ) and Kendall's Rank Correlation Coefficient (τ).

4.4.4 Results and Findings

I report the confusion matrix of terms identified monolingually from the English STs and their Chinese TTs by the classifiers trained on GENIA data (GENIA) and ACL RD-TEC data (ACL) in Table 4.8. Note that these results are reported at the corpus level.

A significant difference of performance is found between the ACL-based classifier and the GENIA-based classifier on both English ST data (z = -1.92, p<$.05$), while the difference between the classifiers on translational data is insignificant (z = -1.29, p>$.05$).

There are 7 real terms in the original ST of XENO data and 22 in the SUIBE data. Because the proposed method works at the corpus level, lots of unwanted n-grams are introduced as noises during the n-gram generation process, which causes a low

		ST			TT		
		precision	**recall**	**F1**	**precision**	**recall**	**F1**
GENIA	XENO	0.04	0.5	0.07	0.17	0.91	0.29
	SUIBE	0.08	0.18	0.22	0.20	0.91	0.33
ACL	XENO	0.14	0.52	0.22	0.27	0.91	0.42
	SUIBE	0.18	0.68	0.28	0.33	0.91	0.48

Table 4.8 Monolingual Terminology Identification on Two Datasets

precision for the classifiers. However, the recalls for both classifiers are high enough to retrieve the majority of the terms in the translations.

On the XENO data, as shown in Table 4.8, in terms of precision, both classifiers perform somewhat consistently on the English source text but display an apparent variation on the Chinese translations instead. I manually analysed their predictions on the termhood of n-grams. Both classifiers have successfully identified 'xenotransplant', 'xenotransplantation','transplant surgeon' as terms, but they failed to identify two terms 'recipient' (受体) and the institute 'America's Food and Drug Administration'. A possible explanation is that they are ignored because they are singletons making it difficult for the statistics-based features to capture the subtlety. Also, the n-gram length threshold is set to be 5, which neglects phrases longer than 5 words, such as 'America's Food and Drug Administration'. For the Chinese translations, all the representative terms, such as 器官 移植 医生 ('transplant surgeon') and 异种 器官 移植 手术 or 异种 器官 移植 ('xenotransplantation'), have been successfully identified.

Meanwhile, ACL-based classifiers manifested a significant improvement of recall. On average, over half of the true terms (22) from the ST could be recognised by the trained classifiers. This is in contrast to the excellent recall on the translation side, as shown in Table 4.8. Term equivalents in the translations for those ST terms that were misclassified can be identified by the trained classifiers (together with a significant proportion of false positives). As I directly apply the classification model trained from different domains, I suspect the domain-shift issue still impacts the classification.

In the end, I use the results from the ACL-based classifier to compute the normalised term count as a quality indicator. All the positive terms identified from the translations by this model are used as 'true terms' that are successfully translated. We must admit that this is rather a naive approach. The assumption behind the approach is that quality translations should roughly contain a proportionally larger finite set of terms. Using the list of 'true terms' identified as reference terms, we can obtain the number of terms contained in each translation through query n-grams in the translations. These numbers are then normalised with the length of the ST and TT respectively as terminology translation precision (normalised by TT length) and recall (normalised by ST length). In the following, we are to see how these numbers correlate with the quality scores.

Dataset	Terms	Human Annotation	Correlation		
			Pearson	Spearman	Kendall Tau
XENO	$\mu = 3.68, SD = 3.45$	UT ($\mu = 24.31, SD = 4.73$)	$r = 0.43, p < 0.01$	$\rho = 0.48, p < 0.0001$	$\tau = 0.37, p < 0.0001$
		TS ($\mu = 16.67, SD = 3.06$)	$r = 0.46, p < 0.01$	$\rho = 0.52, p < 0.0001$	$\tau = 0.39, p < 0.0001$
		IW ($\mu = 17.12, SD = 2.63$)	$r = 0.32, p = 0.02$	$\rho = 0.35, p = 0.01$	$\tau = 0.26, p = 0.01$
		TM ($\mu = 9.79, SD = 1.35$)	$r = 0.36, p = 0.01$	$\rho = 0.39, p < 0.01$	$\tau = 0.31, p < 0.01$
		Total ($\mu = 71.57, SD = 12.41$)	$r = 0.66, p < 0.0001$	$\rho = 0.72, p < 0.0001$	$\tau = 0.55, p < 0.0001$
SUIBE	$\mu = 10.52, SD = 10.19$	Total ($\mu = 87.07, SD = 5.86$)	$r = 0.53, p < 0.001$	$\rho = 0.60, p < 0.001$	$\tau = 0.44, p < 0.0001$

Table 4.9 Correlation between Term occurrences and Translation Quality

#	ST	TT
1	Transplant surgeons work miracles.They take organs from one body and integrate them into another , granting the lucky recipient a longer , better life .	器官 移植 外科 医生 带来 了 奇迹 。他们 将 器官 从 一 个 身体 中 取出 并将 它们 植入 他 者 体内 ， 让 那些 有幸 得到 它们 的 人 活 得 更 长 ， 更 好 。
2	America's food and drug administration has already published draft guidelines for xenotransplantation. The ethics of xenotransplantation are relatively unworrying .	美国 的 食品 药物 管理 机构 已经 出版 了 异种 器官 移植 草案 准则 。这 种 手术 在 伦理 道德 领域 相对而言 ， 不 那么 令 人 担忧 了 。
3	So far attempts to make artificial organs have been disappointing : Nature is hard to mimic. hence the renewed interest in trying to use organs from animals .	到 目前 为止 ， 试图 人工 制造 器官 的 可能性 已经 被 否定 了 ， 毕竟 自然 是 很 难 去 模仿 的 ， 因此 ， 人们 正 将 更 多 的 目光 集中 在 动物 器官 上 。

Table 4.10 Adequately Translated Terms:Term Variation

According to the values of three correlation metrics in Table 4.9, for the XENO dataset, the number of terms identified in both datasets show a positive linear relationship between weak and moderate ($p < 0.01$) with the four subscores. In contrast, the occurrence of terms with the total score (weighted summation of all subscores) goes up beyond moderate ($p < 0.0001$). For the SUIBE data, as it has only one total score for all translations, we could also find a moderate linear relationship between the rightly translated terms in the translations ($p < 0.001$). Despite the fact that two datasets are evaluated by different annotators under different criteria, correlation scores, either Pearson r, Spearman ρ or Kendall's τ all suggest that the number of normalised translated terms does contribute to translation quality on the whole.

However, it is surprising that only a moderate correlation exists (≈ 0.5) between the second subscore (Terminology) and the term occurrence in the translations.

I checked those translations with zero hit of terms but over-strong human quality scores. I found the translation of terminology, semantic adequacy and language fluency indeed present in those translations, see Table 4.10. Typical terms in the specific domain, such as 异种 器官 移植 ('xenotransplantation'), 器官 移植 外科 医生 ('transplant surgeons') and 美国 食物 药物 管理局 ('America 's food and drug administration'), are adequately translated. One thing in common with these translations is that in the translation terms are rendered with a slight variation. For example, in one sample, both 器官 移植 外科 医生 ('transplant surgeons') and 器官 移植 手术 师 ('transplant surgery technician') are used for the same source term 'transplant surgeon'. Both translations are acceptable expressions in Chinese for annotators concerning adequacy and fluency. This term inconsistency or variation may have to do with the reason why such translations are evaluated reasonably high even with few or no term counts by the trained term classifiers. This shows the importance of proper handling of term variation in assessing translation quality.

4.5 Summary

In this chapter, I explored ways of identifying terms from monolingual texts and integrate them into investigating the contribution of terminology to translation quality.

The researcher first reviewed approaches to term recognition. Bilingual term extraction extends further on the basis of monolingual term extraction, making use of various methods and resources, such as parallel and comparable corpora. Then the researcher moves on to propose a supervised learning method to train term classifiers from monolingual data in one domain and extrapolate the models for different languages (English-Chinese) and domains, e.g. from medical to technology. The method has demonstrated reasonably good accuracy on the cross-domain and cross-language data.

To find out how automatically extracted terms impact the translation quality, I then carry out correlation analyses on two small collections of domain-specific translations (xenotransplantation and patent) that are manually evaluated per different criteria. It is found that the number of term frequencies identified automatically has the above-weak linear correlation with the four subscores for the xenotransplantation data. When it comes to the overall final score for both datasets, such correlations increase to be moderate. This study indicates that term occurrence in translation identified this way could be a potential indicator for HTQE.

Chapter 5

Data, Annotation and Translation Error Analysis

5.1 Introduction

In this chapter, I move on to introduce the datasets I use for the QE task, and how quality-indicating features are extracted from them with regard to techniques, tools and resources. In the meantime, I will discuss the method and scheme used for the annotation as well, paying particular attention to translation error analysis of a small proportion of annotated HT and MT data.

In brief, this chapter serves four purposes: **a)** describe the datasets used in the experiments; **b)** introduce whenever possible the techniques, tools and resources involved to prepare the training and testing data; **c)** carry out a descriptive analysis of the engineered features with respect to translation quality; **d)** compare translation errors in HTs and MTs.

For researchers in HTQE, one of the ways is to understand the relationship between many variables, including but not limited to linguistic systems, domain and register variation, translators' decision making and task specifications, and the implications these variables may have on the quality of translation products and our perception of translation quality itself. Much research has examined quality evaluation from theoretical perspectives (House, 2014), estimating (machine) translation quality with reference-based metrics (e.g. BLEU, NIST, RIBES) and training predictive models with linguistic (e.g. POS tags, dependency relations) and non-linguistic features (shallow features and system decoding features). Research in MTQE has discussed the effectiveness of a set of features (Avramidis and Popovic, 2013b; Luong et al., 2013; Shah et al., 2013) during the feature selection process for estimating pseudo human scores (e.g. HTER). However, research on pairwise associations of quality indicators with human quality scores (labels) has been limited (Yuan et al., 2016).

I argue that the importance of individual quality indicators should not be undermined, particularly for the reason of interpretable results in HTQE, while the model

accuracy is prioritised. Therefore, the **main contribution** of this chapter lies in two aspects:

- statistical analysis of the distribution of translation errors that helps uncover the fundamental relationship between translation quality and translation types (e.g. machine translation and human translation), text types and text domains.
- pairwise association analysis of individual quality features with different types of manually annotated quality scores (7 in this study) based on a large feature set (360 at the document level).

5.2 Data

The translational data used for the training and testing in this study come from the published Parallel Corpus of Chinese EFL Learners (Wen and Wang, 2008). The corpus consists of two components, namely the written translations and interpretation transcripts by trainee translators, who are all third-year or fourth-year English majors. The written translation part is composed of 2385 translated documents for 6 STs in different domains. In this study, 457 of them are selected as training and testing data for QE at the document level. Among these 457 translations, every 50 documents are taken from the first group (each source text has three batches of translations by trainee translators from different universities) of translations to 6 STs, using stratified sampling. The other 157 texts are taken from the second and third group ($\mathrm{ST1\_2}$ and $\mathrm{ST1\_3}$) translations to the first ST as an additional data. In this sense, all translations to the first ST are included. Table 5.1 below lists the basic statistics of the written translations annotated. Note that these statistics are based on the tokenised English source text and the segmented Chinese. In the seventh column of this table, the first number indicates the number of documents selected and included in the dataset, and the differences in $\mathrm{ST1}_2$ and $\mathrm{ST1}_3$ are due to discarding some empty documents. Examples of the STs and trainee translations are provided in the Appendix A.

5.3 Annotation

In this section, I describe the tools and resources used to process the STs and TTs and the quality annotation scheme I adopt to assign scores to TTs. In addition, details of the feature extraction method are also discussed.

5.3.1 Tools and Resources

As discussed in Section 3.2, most monolingual features I selected or designed are dependent on linguistic analysis of the STs and TTs at the lexical, semantic and

ST	Domain	Topic	Statistics					
			ST		TT			
			# Sent.	# Words	Category	# Doc.	Avg. # sent.	Avg. # Words
ST1	Science fiction	Insects	11	261	ST1_01	50/137	10.93	282.10
					ST1_02	56/56	9.89	244.88
					ST1_03	101/102	11.00	260.08
ST2	Scoial life	Marriage	15	259	ST2_01	50/153	14.65	251.10
					ST2_02	115	14.98	258.23
					ST2_03	141	14.99	259.14
ST3	Sports	Walking	13	289	ST3_01	50/254	12.95	296.43
					ST3_02	93	13.00	290.43
					ST3_03	198	12.34	267.41
ST4	Short story	Perseverance	15	313	ST4_01	50/122	14.83	342.57
					ST4_02	154	14.98	345.53
					ST4_03	112	14.87	324.90
ST5	Literature	Essayist	5	229	ST5_01	50/103	5.00	220.31
					ST5_02	117	4.99	216.09
					ST5_03	109	4.93	209.98
ST6	Science	xenotransplantation	13	266	ST6_01	50/136	12.85	307.71
					ST6_02	142	12.99	303.08
					ST6_03	141	12.72	291.30

Table 5.1 Basic Statistics of Source and Translation Data

discourse level. In the following, I briefly summarise the tools and resources I have relied on to extract the corresponding features.

- POS Tagging and Parsing
  First, all English STs and Chinese translations are tokenised or segmented. For the English STs, I use the scripts coming with Moses decoder[1]. The Chinese TTs are segmented with Jieba[2]. The tokenised or segmented STs and TTs are further analysed with Stanford Parser (Klein and Manning, 2003; Levy and Manning, 2003) for POS tags, constituencies and dependencies.

- Semantic Role Labelling
  The distribution of semantic role labels in both STs and TTs help disclose the underlying relationship between the main verbs and other participants in the clause. It is a rather abstract way of semantic representation. Comparing them in ST and TT lends insight to the quality itself. As mentioned earlier in Section 3.2, English and Chinese language often demonstrate systemic variation in choosing SR labels. I chose PathLSTM (Roth and Lapata, 2016) for English STs, and the Language Technology Platform API (Che et al., 2010) for Chinese TTs.

- Machine Translation Systems
  To compute the pseudo references and back translation similarity features, I use three commercial machine translation systems, namely Google Translate

[1] https://github.com/moses-smt/mosesdecoder/tree/master/scripts/tokenizer
[2] https://github.com/fxsjy/jieba

Window Size	Dynamic Window	Sub-sampling	Low-Frequency Word	Iteration	Negative Sampling
5	Yes	1e-5	10	5	5

Table 5.2 Hyper-parameters for training Chinese embeddings

Corpus	Size	Tokens (Million)	Vocabulary (Million)	Description
Chinese Wikipedia	≈ 1.3G	≈ 230	≈ 2.2	https://dumps.wikimedia.org/

Table 5.3 Detailed Information of the Chinese Wikipedia Corpus

[3], Bing Translator [4] and Yandex [5] who offer the API access. I use these MT systems to translate the STs into TTs, and TTs back into pseudo STs.

- Word Representation
  Word vectors such as GloVe (Pennington et al., 2014) have proven to capture fine-grained semantic and syntactic information (Mikolov et al., 2013c). In this study, I use the pretrained monolingual embeddings and bilingual embeddings for computing document coherence (adjacent sentence cosine distance, refer to section 3.2.1.5 for more detailed explanation). For English, I use the 200 dimension Glove vector[6]. For Chinese, I trained vectors of the same dimension size on Chinese Wikipedia (Table 5.3) using the skip-gram model with negative sampling implemented in Gensim (Řehůřek and Sojka, 2010). Table 5.2 details the basic settings. For computing the STs and TTs similarity directly from word representations, bilingual embedding with a dictionary-based mapping technique (Artetxe et al., 2017) is employed to normalise monolingual source and target embeddings with only some manually prepared English-Chinese lexicon. On its basis, sentences in STs and TTs are represented as the average of word embeddings for words in them.

- Parallel and Comparable Corpora
  In order to train word and phrase alignment models and the bilingual term extraction model, I use a collection of parallel and comparable corpora. As regards the parallel corpora, I use the English Chinese parallel UM corpus of mixed domains (Tian et al., 2014), as listed in Table 5.4. It is a multi-domain and balanced parallel corpus covering several topics and text genres, including education, law, microblogs, news, science, spoken, subtitles and thesis.

  Details of comparable corpora have been provided in Section 4.3.2.2 for crosslingual terminology extraction.

- Alignment Software
  For word and phrase alignment in this study, I use the tool Pialign (Neubig et al., 2011).

[3]https://translate.google.com/, accessed on December 22, 2017
[4]https://www.bing.com/translator, accessed on December 22, 2017
[5]https://translate.yandex.com/, accessed on December 22, 2017
[6]http://nlp.stanford.edu/data/wordvecs/glove.6B.zip

Domains	Languages	Tokens	Avg. Len.	Vocabulary	Sent.
News	English	8,646,174	19.21	274,546	45,000
	Chinese	15,277,414	33.95	47,902	
Spoken	English	1,836,670	8.35	107,923	220,000
	Chinese	3,033,052	13.79	9,011	
Laws	English	5,926,316	26.94	66,330	220,000
	Chinese	8,783,941	39.93	14,723	
Thesis	English	5,962,590	19.88	378,679	300,000
	Chinese	10,514,430	35.05	149,110	
Education	English	8,401,095	18.67	293,595	450,000
	Chinese	13,749,570	30.56	38,663	
Science	English	598,050	2.22	115,968	270,000
	Chinese	1,527,849	5.66	8,927	
Subtitles	English	2,299,742	7.67	101,423	300,000
	Chinese	3,818,490	12.73	13,854	
Microblog	English	72,144	14.43	12,083	5,000
	Chinese	125,415	25.08	3,525	
Total	English	33,742,781	13.29	832,518	2,215,000
	Chinese	56,830,161	22.51	209,729	

Table 5.4 Statistics of UM Parallel Corpora

Bayesian-based phrase alignment as proposed in Neubig et al. (2011) is an unsupervised model for joint phrase alignment and extraction using non-parametric Bayesian methods and inversion transduction grammars (ITGs) (Wu, 1997).

As noted by DeNero and Klein (2008), phrase alignments generated through the flat model are not optimal, as there are only minimal phrases memorised by the model and it has to be combined with heuristic phrase extraction to combine exhaustively adjacent phrases permitted by the word alignment (Och et al., 1999). A hierarchical ITG model relies on the Pitman-Yor process (Pitman and Yor, 1997) to directly use probabilities of the model as a replacement for the phrase table generated by heuristic techniques, e.g. intersection, grow-diag in Giza++ (Och and Ney, 2003).

I build a list of good quality phrase alignment pairs from a bilingual corpus of professional translations to measure the degree of adequacy and fluency of trainee translations. I extracted phrases up to a maximum length of 4. The reason I did not go up to more extended sequences of phrases is that longer alignments (>4 words) are very rare in students' translations. I include one word as phrases for the consideration of one-to-many and many-to-one alignment (i.e. many phrases are translations of one source word, and many one words are translations of source phrases), and the avoidance of an addition process of word alignment. Having used Pialign to train on the UM Parallel Corpora, I eventually have 9.63 million pairs of phrasal alignments (including word alignments) without exclusion and ≈ 5.72 after filtering per direct translation probability (DTP) and inverse translation probability (ITP) (DTP and ITP $>= 0.01$) (See Table 5.5).

Length	Counts	Alignment Types	Counts
1-word	648,611	one-to-one	374,840
2-word	1,835,261	one-to-many	273,771
3-word	1,889,009	many-to-one	362,849
4-word	1,344,276	many-to-many	4,705,697
Total	5,717,157	**Total**	5,717,157

Table 5.5 Length and Type Distribution of Alignments

Alignment	Top1000(%)	Bottom1000(%)
word alignments	96.3	26.9
phrase alignments (2-more words)	98.7	97.2

Table 5.6 Alignment Accuracy (threshold DTP $>= 0.2$)

I compared the top 1000 and bottom 1000 entries for single word alignments and phrasal alignments (2 to more words). The researcher manually checks the two lists of selected aligned words and phrases to see if they are translations of each other, and for phrase alignments, I also consider 'near translations', i.e. those partial alignments with additional or fewer words. For instance, 交通罪行 的 ('traffic offence about') for traffic offences in Table 5.7 is an acceptable phrase alignment. I compute the accuracy of alignment for both words and phrases this way to compare the quality of alignments.

Longer phrase alignments from the training corpus are generally more accurate than shorter ones. To be more specific, the accuracy of phrase alignments remains stable even though their direct translation probabilities decrease. In order to evaluate the validity of entries in the extracted list of phrase alignments, I set the DTP threshold for all alignments to be 0.2 and then sort them per their DTP values and discard all non-English-Chinese pairs, i.e. Null alignments, punctuations, symbols, strings alignments and English-English alignments. That is to say, the evaluation is based on alignment pairs consisting of words only. See Table 5.6 for details.

I conjecture that the reason for many false matches in one-word alignments is word segmentation and word association in the near context. For instance, I found both 无用功 and 拖垮 matched by 'unproductive'. 无用功 ('unproductive work') should be matched by two English words instead of one 'unproductive', and 拖垮 ('drag down') should be the cause of 'unproductive work' but it is mismatched as an alignment because it occurs within the near context, and also due to the high probabilistic nature of the alignment algorithm.

Nevertheless, as Table 5.6 shows, these phrase alignments, particularly those longer than one word, when selected as bilingual correspondences with reasonably good accuracy, can be readily usable.

Length	English	Chinese	DTP	Gloss
1-word	offences	所犯	0.33	committed by
		犯罪行为	0.29	crimes
		而犯	0.44	commit
		系由该	1.00	This is committed by
		犯罪	0.22	commit a crime
2-word	traffic offences	违犯 交通法规	0.91	violation of traffic regulations
		交通 罪行 的	1.00	traffic offence (about)
		交通 罪行	1.00	traffic offence
3-word	international drug trafficking	国际 贩毒	0.39	international drug trafficking
		国际 毒品 贩运	0.86	international drug trafficking
		国际 药物 贩运	1.00	international medicine trafficking

Table 5.7 Illustration of Alignments

In preparing the dictionary of bilingual phrase alignments for query their occurrences in trainees' translations, four-word alignments are eventually discarded because of their extremely low frequencies in the target translations. Table 5.7 illustrates the alignments.

Note that while DTP may be a useful signal of alignment certainty, we cannot take it for granted that lower probabilities nullify the legitimate translation equivalents. Many pairs of very low probabilities are valid translations to each other. For instance, 报价和目录 (gloss:Quote and directory) and 'quotation and catalogues' are aligned at a probability of 0.08, but they are clearly valid alignments. This fact again explains why in Table 5.6 there is no significant difference for longer phrase alignments ranked by DTPs (top and bottom). Also, this shows that we need to keep good coverage of phrase alignments. For this study, I set the threshold of direct translation probability at 0.02. This cutting-off value eventually allows us to have 3.5 millions pairs of the word and phrase alignments, a much slimmer list of phrase table, as illustrated in Table 5.8. In the Table, the third and fourth columns are direct (conditional) translation probabilities (DTP, i.e. translation probability from English to Chinese) and inverse translation probabilities (ITP, i.e. translation probability from Chinese to English).

With this acquired probabilistic dictionary of word and phrase alignments, each 1- word, 2-word and 3-word lexical unit in the TT is then queried against each trainee translation to find the matches between the corresponding units. These matches are then normalised as quality features.

Source Segments	Target Semgents	DTP	ITP
tax credits	税额 减免	0.37	0.25
	税款 抵减	0.52	0.34
	税收 减免	0.08	0.05
	的 税收 优惠	0.90	0.05
	税收 优惠	0.05	0.05
	税额 抵免	1.00	0.05
	税收 抵免	0.59	0.17
and tax credits	或 减免	0.50	0.50
	和 税 减免	1.00	0.50
as tax credits	诸如 税收 扣减	1.00	1.00
continuous tax credits	持续性 税收 扣除	1.00	1.00
for tax credits	享受 税额 减免	1.00	1.00
investment tax credits	投资 税额 减免	1.00	1.00
in tax credits	税收 抵免	0.02	1.00
production tax credits	生产 税额 减免	0.33	1.00

Table 5.8 An Excerpt of Alignments

5.3.2 Quality Annotation

Quality scores annotated by human raters are the learning goal of the proposed approach. In MTQE, translation quality are evaluated in terms of adequacy and fluency (Callison-Burch et al., 2006), post-editing effort, e.g. HTER, post-editing time (Bojar et al., 2014, 2015, 2016a).

I chose the American Translators Association (ATA) Certification Programme Rubric for Grading (hereinafter ATA rubric for short) (ATA, 2011) as the annotation guideline. The rubric was used in the ATA translation qualification certification test, allowing for a more systematic and holistic grading (Angelelli and Jacobson, 2009b). More importantly, it assesses a few subcomponents of translation quality which suits our need for fine-grained quality estimation. Each subcomponent corresponds to a different aspect of translation quality. In the following, I briefly describe the ATA rubric as our annotation guideline.

ATA rubric intends to evaluate test takers' performance holistically in four subcomponents:

- **usefulness** indicates to what extent the translation is usable and the meaning of the source text has been conveyed.
- **terminology** is more about the lexical appropriateness of the translation in terms of terminology, register and style.
- **idiomatic writing** emphasises the smoothness of the translation in the target language.

- **target mechanics** requires to follow target language norms.

Each component is further categorised into five levels to represent the quality in that dimension:

- **standard**
  The target text would require little if any editing.
- **strong**
  The target text requires only minimal work to be published or used.
- **acceptable**
  The target text could be used with some post-editing work.
- **deficient**
  The target text requires extensive editing work to be published or used.
- **minimal**
  The target text has to be retranslated.

The ideal performance for each dimension is defined in the 'standard' row against which the annotators mark the translation accordingly. For instance, the standard criterion for '**usefulness/transfer**' is specified as:

> The translated text is fully usable for the purpose specified in the Translation Instructions. The meaning and sense of the source text have been fully and appropriately transferred to the translated text.

In contrast, the worst performance for each dimension is labelled as 'minimal', which implies it would be more economical to have the text retranslated. A'**minimal**' level of ' bf usefulness/transfer' would suggest:

> Translated text transfers meaning in a manner inconsistent with the Translation instructions. Translation contains frequent and/or serious transfer errors that obscure or change meaning.

More detailed explanations of the grading components and different levels are provided in Appendix B. I assign different weights to each component. More specifically, I assign more weights to components related to adequacy (e.g. usefulness, terminology) over components related to fluency (e.g. idiomatic writing, target mechanics), as adequacy generally has to do with completeness and meaning transfer of the translation while fluency is largely associated with the target language conventions. Thus, we have four component scores which are further collapsed into adequacy and fluency, two classic indexes for quality evaluation, i.e. the first two scores attribute to adequacy and the other two for fluency. For adequacy and fluency, I use the definition and evaluation scale created by the Advanced Research Projects Agency (ARPA) (Church and Hovy, 1993; White, 1994). Adequacy measures how

much source information is preserved in the translation, related to the correctness of the translation. Adequacy is often calculated on a Likert-like scale corresponding to the five ATA categories as shown in Table 5.9. In contrast, fluency can be defined as how fluent the translation is regarding target language quality. Fluency is calculated on the same rank scale as adequacy in Table 5.10.

Score	ATA Categories	Definition
1	minimal	None of the meaning is preserved
2	deficient	Little of the meaning is preserved
3	acceptable	Much of the meaning is preserved
4	strong	Most of the meaning is preserved
5	standard	All the meaning is preserved

Table 5.9 Adequacy Evaluation Scale

Score	ATA Categories	Definition
1	minimal	Incomprehensible target language
2	deficient	Disfluent target language
3	acceptable	Non-native kind of target language
4	strong	Good quality target language
5	standard	Flawless target language

Table 5.10 Fluency Evaluation Scale

In this sense, for each translation we have seven scores for each translation, namely, *usefulness* (UT), *terminology* (**TS**), *idiomatic writing* (**IW**), *target mechanics* (**TM**), *adequacy* (**AD**), *fluency* (**FL**) and *total* (**TO**) amounting to 100 points. In other words, 35 points for **UT** 25 points for **TS**, 25 points for **IW** and 15 points for **TM**, 60 points for **AD**, 40 points for **FL**, and 100 points for **TO**. I propose a range finder to help the annotators score the translations with flexibility, as shown in Table 5.11.

Grades	Usefulnes	Terminology	Idiomatic Writing	Target Mechanics	Adequacy	Fluency	Total
Standard	29-35	21-25	21-25	13-15	50-60	34-40	84-100
Strong	22-28	16-20	16-20	10-12	38-49	26-33	64 -83
Acceptable	15-21	11-15	11-15	7-9	26-37	18-25	44-63
Deficient	8-14	6-10	6-10	4-6	14-25	10-17	24-43
Minimal	1-7	1-5	1-5	1-3	1-13	1-9	1-23

Table 5.11 Range Finders for Different Grades of Translation

All selected translations (457 documents and 3569 sentences) are then annotated by two annotators per the guideline for later training and testing the models to predict the quality of trainee (learner) translations. Annotators are encouraged to score fluency separately from adequacy in their capacity. The inter-annotator

agreements (Krippendorff's Alpha) for the first four components are reported below in Table 5.12

		Annotator A	Annotator B	
		μ		α
Document-level	Usefulness	23.02	23.15	0.89
	Terminology	16.99	17.28	0.85
	Idiomatic writing	17.85	17.92	0.74
	Target meachnics	9.79	9.75	0.96
Sentence-level	Usefulness	22.06	22.26	0.96
	Terminology	16.6	16.63	0.96
	Idiomatic writing	19.03	19.23	0.74
	Target meachnics	10.67	10.86	0.89

Table 5.12 Inter-annotator Agreement on English-Chinese HTs

Finally, for every translation document or sentence, the arithmetic means of scores by two annotators are obtained as the final score for each component.

5.3.3 Error Annotation

As my approach to HTQE and the reference-free MTQE are essentially the same in that both methods are supervised learning on the basis of annotated data (manually or automatically) with different features, In this sense, it is necessary to look into the difference of MT translations and human translations. As far as quality is concerned, analysis of the generated translations is beneficial for identifying the main problems and making researchers' work more focused. To compare trainee translation quality with MT I briefly discuss frameworks of translations errors used for analysing MTs and HTs, and I propose an adapted error typology to annotate the data.

5.3.3.1 MT Error Typology

Font-Llitjós et al. (2005) propose a framework for identifying and correcting rules semi-automatically in order to improve translation coverage and quality, in which they define a preliminary MT error typology. There are 5 main classes: missing word, extra word, wrong word order, incorrect word, and wrong agreement. Under the wrong word order subclass, there are subclasses of errors including local versus long distance, word versus phrase, and word change. Incorrect word errors are further divided into sense, form, selection restriction, and idiom errors. Missing constraint and extra constraint are representative of the wrong agreement. These errors are used by bilingual speakers to identify problematic MT translations during the correction process to instantiate the error information that triggers a particular correction in the system.

Vilar et al. (2006) propose a taxonomy of error types, such as unknown word, incorrect word form and long-range word order. The classification scheme is an extension of the framework presented in Font-Llitjós et al. (2005). Bojar (2011) then uses the same error scheme for error analysis of English-Czech MTs. This scheme consists of a hierarchy structure of errors, as shown in the Figure 5.1. There are five main classes on the first level: missing words, word order, incorrect words, unknown words, and punctuation errors.

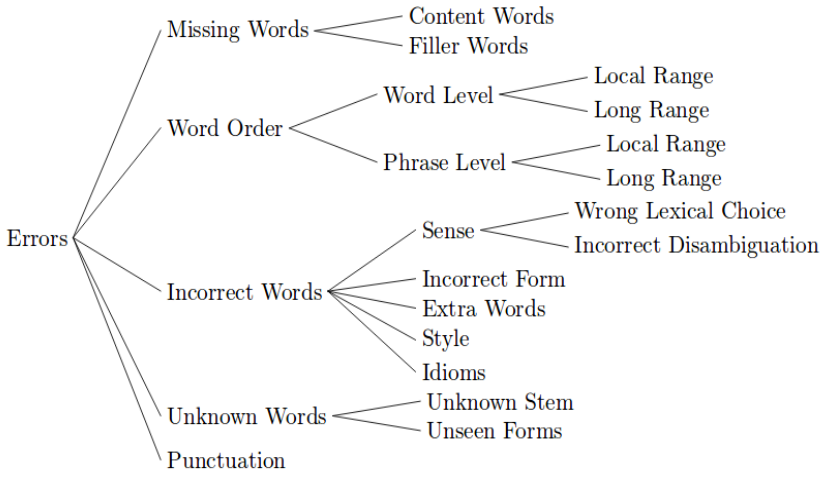


Figure 5.1 Vilar et al (2006) error categories

The first type of errors is produced when some word in the translation is missing, by which the meaning and grammar of the translated sentence have been affected. The word order category concerns the movement of individual and sequences of words to form a right translation. The subcategories within incorrect words refer to translation errors produced by the system with respect to meaning, form (e.g. tense, POS), style (e.g. repetition in a near context), and idioms (e.g. translating an idiom as regular text). Unknown words are unknown and unseen word forms, and punctuation errors are for punctuations breaking the corresponding punctuation rules between language pairs.

Popovic et al. (2006) and Popović and Ney (2007) carry out analysis of reordering and inflectional errors due to syntactic differences between two languages, and propose a method to decompose errors over different POS classes. They use the morpho-syntactic information in combination with the automatic evaluation measures WER and Position Independent Word Error (PER). Syntactic errors have been measured by the relative difference between WER and PER for nouns, verbs and adjectives for the ST-TT pair. Inflectional errors are presented as the relative difference between full form PER and base form PER for different word classes. Their method is further refined and compared with human annotators in (Popović and Burchardt, 2011). For instance, an inflection error is defined as whenever the full form is a reference PER or hypothesis PER, but the base form is correct.

Type	Subtype
morphology	verbal nominal
missing word	function word content word
word sense error	
word order error	short range long range
punctuation	
spelling	
superfluous word	function word content word
capitalization	
untranslated word	medical term proper name other
pragmatic	
diacritics	
other	

Table 5.13 Kirchhoff et al. (2012) error categories

Label	Description
ER	missing words extra words wrong word word order
Ling	orthography semantics syntax
GF	grammatical words functional words
Form	morphological categories
POS+	part of speech punctuation
FA	fluency adequacy both neither
Reo	cause of reordering
Index	position of error
other	other categories

Table 5.14 Stymne and Ahrenberg (2012) error categories

Farrús et al. (2010) present a coarse error scheme of five main classes, i.e. orthographic, morphological, lexical, semantic and syntactic, for comparing MT systems. They report that certain errors, e.g. lexical and semantic, are perceptually more important than other errors for human evaluators. Comelles et al. (2012) use a similar scheme as a basis of linguistic features to develop automated MT evaluation metrics.

Federico et al. (2014) use another similar typology which contains

- morphological errors
- lexical choice
- additions
- omissions
- casing and punctuation
- reordering errors
- too many errors

for annotating MTs from English into Arabic, Chinese and Russian. Annotators mark MT errors and assign an overall quality score with the evaluation guidelines and the annotation tool (Girardi et al., 2014). On its basis, mixed-effects modelling

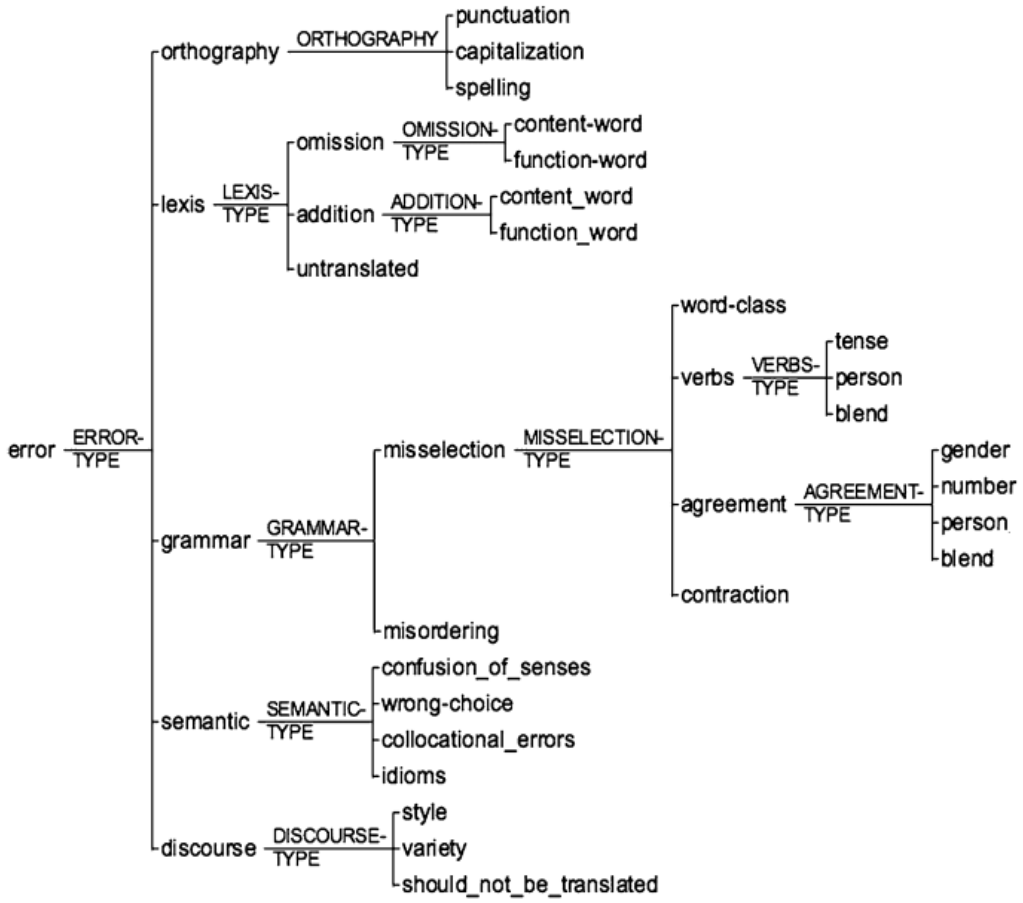


Figure 5.2 Costa et al. (2015) error categories

(Baayen et al., 2008) is then employed to investigate the impact of particular error types and their interaction on the overall quality score.

Using a set of error (sub)types as in Table 5.13, Kirchhoff et al. (2012) apply conjoint analysis (Green and Srinivasan, 1978), a method to determine what combination of attributes is most influential on respondent choice or decision making, to the annotated MT outputs and examine user preferences of different error types. They have found word order and word sense errors, followed by morphological errors, are most influential. Their findings resemble Federico et al.'s which maintains missing words also have the largest correlation with translation quality.

Stymne and Ahrenberg (2012) investigate the inter-annotator agreement of translation error classification, using a hierarchical error typology (See Table 5.14). Two annotators assign error classes in two rounds, with or without some general guidelines. They have found the inter-annotator agreements for the detailed error items and the simple typology in the round with guidelines increased significantly, when compared to those without guidelines in the second round.

Costa et al. (2015) propose an extensive error taxonomy of translation errors (See Figure 5.2) targeting morphologically richer Romance languages, comparing four systems translating from English to European Portuguese. Their study has shown lexical (e.g. untranslated) and grammar errors (e.g. misordering) impact most

the ranking of MTs at the sentence level, followed by semantic errors (confusion of senses and wrong choice).

5.3.3.2 HT Error Typoplogy

Manual examination of translation errors has been common for HT evaluation. LISA QA metric[7] from the Localization Industry Standard Association categorises translation errors and their severities into minor, major, and critical. Language (e.g. mistranslation, accuracy), formatting (e.g. layout, graphics) and functionality testing are the main categories. SAE J2450 QA model[8] contains 7 types of errors in the minor and serious grades. ATA standardised error marking ATA (2017) uses a set of 21 error categories[9] in their certification examinations. ATA errors are roughly in three groups dealing with the form (e.g. illegibility), meaning transfer (e.g. literalness) and language mechanics (e.g. spelling). Instead of the verbal categories of seriousness, ATA error framework adopts a numerical scale (1-16) indicating different degrees of negative impacts, which are guided by a flowchart for error grading[10].

Solano-Flores et al. (2009) identified ten test translation error dimensions: style, format, conventions, grammar and syntax, semantics, register, information (e.g. changes in the way numbers are written), construct (e.g. omission, insertion of technical terms), curriculum (e.g. testing content is not covered in the curriculum), and origin (e.g. source flaws carried over to the target). Each dimension comprises several translation error types. For instance, the style dimension includes incorrect use of accents, incorrect use of uppercase letters, incorrect use of lowercase letters, subject-verb inconsistency, spelling mistakes, incorrect punctuations. These error dimensions can be grouped into three broad categories: language, content, and item design.

Multidimensional Quality Metrics (MQM) (Lommel et al., 2014b) is a unified paradigm for HT and MT quality assessment, which is designed as a superset of the core 'issue types' identified in previous measures and tools (e.g. LISA QA Model, SAE 2450). MQM categorises translation issues into five main classes (Lommel et al., 2013):

- **Fluency** which is related to the language itself of the translation, regardless of its status as a translation.
- **Accuracy** which is related to how well the content of the target text is conveyed.

[7] http://producthelp.sdl.com/SDL_TMS_2011/en/Creating_and_Maintaining_Organizations/Managing_QA_Models/LISA_QA_Model.htm

[8] http://producthelp.sdl.com/SDL_TMS_2011/en/Creating_and_Maintaining_Organizations/Managing_QA_Models/SAE_J2450_QA_Model.htm

[9] https://atanet.org/certification/Framework_2017.pdf

[10] https://atanet.org/certification/aboutexams_flowchart.pdf

- **Verity** which concerns whether the translation meets the real world requirement.
- **Design** which is concerned with the formatting and style of the translation.
- **Internationalization** which is concerned with the internationalisation of the content.

The complete catalogue of issue types contains more than 108 translation errors. A core list of issue types is downsized to 3 main classes: accuracy, verity, and fluency, and 18 subcategories. The MQM is then harmonised with TAUS Dynamic Quality Framework (DQF) to be the DQF-MQM error typology[11] that offers a template to allow users to categorise and count translation errors segment-by-segment in HT and MT texts. The typology comprises 8 major types and 33 subtypes of translation errors.

5.3.3.3 An Adapted Error Typoplogy

I seek not to estimate translation quality by translation errors. Instead, I indent to understand how translation quality variation is embodied in the distribution of translation errors. In other words, the patterns characteristic of error distribution for both MT and HT are more interesting to us. I adapted the DQF-MQM to annotate the translations since some sub-issue types do not fully apply to our dataset. The reason I select this error framework as the basis for the annotation is that it is explicitly designed for describing both MT and HT quality[12]. Though I use the ATA rubric to annotate translation quality in my study, I do not apply its error typology to the error annotation as it is designed for HT certification. Error types which are not adopted in the annotation include over-translation, improper exact TM match, grammatical register, inconsistency, link/cross-reference, character encoding, inconsistent with termbase, inconsistent use of terminology, all errors under style, design and locale convention, and cultural-specific reference. For trainees' written translations, error types in design, such as markup, local formatting, and errors in locale convention, such as address format, date format, are no longer applicable. In the meantime, I use the main class of terminology to replace all subtypes of errors for the sake of creating less confusion for the annotators. The final list of error types used for annotating the data is included below:

- **mistranslation** that the target content does not accurately represent the source content.
- **omission** that content present in the source is missing from the translation.
- **awkward** that a text is written with an awkward style.

[11] https://www.taus.net/quality-dashboard-lp#dqf-mqm_error_typology
[12] http://info.taus.net/dqf-mqm-error-typology-templ

- **punctuation** that punctuation is misused for the target language.
- **undertranslation** that the target text is less specific than the source text.
- **unidiomatic** that the content is semantically correct but not as natural as native target texts.
- **grammar** that the target text manifests grammatical and/or syntactic fallacies.
- **addition** that the target text includes content not present in the source.
- **spelling** that the target text has deficient written forms, e.g. spelling error, made-up words.
- **terminology** that a domain-specific word is translated into an inappropriate term or a non-term.
- **untranslated** that content that should have been translated has been left untranslated.

This list is a somewhat slimmer version of error types. However, it gives the annotators flexibility and causes less confusion, given that sometimes translation errors are not mutually exclusive (Vilar et al., 2006). That is, one type of error often leads to another potential error to occur in the near context. For example, a mistranslation may cause the phrasal and/or clausal disorder. I take a flexible approach to the error annotation by providing only the definitions above and allowing the annotators to decide what best fits in the error typology, though It may be criticised for the lack of scientific rigidity caused by this flexibility. However, taking a more positive view, I argue that equipped with sufficient translation training and bilingual competence, annotators do not necessarily undermine the annotation if we accept that full inter-annotator (rater) and intra-annotator (rater) agreement are implausible, and the varying subjectivity between and within annotators could in some sense reflect the natural responses to translations.

In contrast to the afore-mentioned ATA error framework, the error categories in our list have broader coverage, despite that they are largely overlapped. More specifically, addition, terminology, grammar, punctuation and spelling are all included in the adapted list and the ATA error framework. However, grammar and punctuation in the adapted list have broader meanings. For instance, grammar in our list is equivalent to grammar, syntax, word forms and usage in the ATA framework, judging by their definitions. In terms of application, a narrowed-down list of errors is clearly more friendly to annotators.

To compare the error distribution in MTs and HTs, I translate the six STs using 7 commercial MT systems (Bing Translator[13], Google Translate[14], Jinshan[15],

[13]https://www.bing.com/translator/
[14]https://translate.google.com/
[15]http://fy.iciba.com/

PROMPT[16], SDL[17], Systran[18], Baidu[19]), and at the same time I randomly select 7 HTs of each source text to form a comparable corpus of MTs and HTs. I score the collection of MTs per the same scoring criteria for HT annotation. Table 5.15 details the inter-annotator reliability for the 42 MT translations of the 6 STs. All the MTs and HTs in this section are annotated in accordance with the adapted error typology presented above (μ stands for mean score andα is the inter-annotator agreement.).

		Annotator A	**Annotator B**	
		μ		α
Document-level	Usefulness	12.61	11.9	0.97
	Terminology	8.9	8.5	0.97
	Idiomatic writing	8.5	8.6	0.94
	Target meachnics	4.3	4.3	0.89

Table 5.15 Inter-annotator Agreement on English-Chinese MTs

5.4 Exploratory Data Analysis

On the basis of the above-mentioned annotation, I carry out exploratory statistical analyses on the data I have collected. In the following section, I report the results from the Principal Component Analysis (PCA) (Abdi and Williams, 2010) of the error distribution in the MT and HT data, and from a correlation analysis of features I extracted in Chapter 3 and Chapter 4 with quality scores.

I am looking for the dominant patterns of error distribution across two translation types, and the interaction of those patterns with other non-quantitative variables, such as source text types and topics (See Table 5.1). Therefore, we select the PCA method. It is a powerful analytical method to reduce the data dimension and project them onto the principal components that explain the most variance of the whole data structure.

Figure 5.3 depicts the different types of errors and their frequencies in HTs and MTs. The $\mathrm{x}-$axis of these bar charts are the raw frequencies of error types in individual translations, and $\mathrm{y}-$axis represents the counts of different error frequencies in them. As is shown in the figure, translation errors are more common among mistranslation, omission, awkward and unidiomatic. It is also noteworthy that certain error types, such as grammar and untranslated are uniquely present in MT translations. These patterns conform to our knowledge since mistranslation, omission and awkward are indeed typical of either human or machine, and machine

[16]http://www.online-translator.com/
[17]https://www.freetranslation.com/
[18]http://www.systransoft.com/lp/free-online-translation/
[19]http://translate.baidu.com/

translations are known to be problematic in grammar and dropping content that should have been translated. Refer to the following examples for better illustration.

Example: MT-Grammar Error

from the top of the mountain , sloping for several acres across folds and valleys were rivers of daffodils in radiant bloom .

从 山顶 开始 ， 倾斜 几 英 亩 [awkward] 的 褶皱 [mistranslation] 和 山谷 是 水仙 花 盛开 的 水仙花 [grammar]

gloss: from top of mountain starting , slope several acres folds and valleys are daffodils in blossom daffodils.

Example: HT-Grammar Error

people already kill pigs both for food and for sport ; killing them to save a human life seems , if anything , easier to justify. however , the science of xenotransplantation is much less straightforward .

人们 为了 食物 和 运动 的 目的 而 杀 了 很多 猪 。 但是 若 任何 事 都 可以 轻易 地 使之 合理化[mistranslation] ， 人们 杀猪 而 为 自身 的 生存 也 是 合理 合理 的[grammar] 。 况且 ， 异种 器官 移植 的 科学 也 变得 简单 ， 易懂 了[mistranslation]

gloss: people for food and sports purpose to kill many pigs . but if anything can be easy to be justified , people kill pigs for their existence too is reasonable . and, xenotransplantation science of too became easier , more understandable

Example: MT-Omission

bees , wasps , ants and termites have intricate societies in which different members are specialized for foraging , defense and reproduction .

蜜蜂 、 黄蜂 、 蚂蚁和 白蚁 有 复杂 的 社会 [omission]不同 成员 觅食 是 专用 于 、 国防 和 复制 [mistranslation]。

gloss: bees , wasps , ants and termites have complex societies different members looking for food is specialized for , defence and copy .

Example: HT-Omission

in Europe and America , herds of pigs are being specially bred and genetically engineered for organ donation .

在 欧洲 和 美国 为 器官 捐赠 饲养 出 了[mistranslation] 成群 的 受过 特殊 饲养 的 猪[omission] 。

gloss: in Europe and America for organ donation have kept herds of been specially bred pigs .

The above four examples (2 HTs and 2 MTs) contain 2 instances of omission and 2 instances of grammar errors. In the first example, 水仙花 盛开 的 水仙花

('daffodils in blossom daffodils') is ungrammatical due to the MT system does not know to associate the 'slope' with 'daffodils' and give it a more idiomatic translation 绵延 ('stretches'), in addition to 倾斜 几 英 亩 ('slope several acres') that reads very awkward due to the failure to translate the metaphoric 'rivers of daffodils'. In the fourth example, 饲养 出 了 ('have kept') mistranslated the present progressive tense 'being specially bred', in addition to the 受过 特殊 饲养 的 猪 ('specially bred pigs') that has omitted the modifier 'genetically engineered'. Other two examples contain the similar errors of mistranslation and omission.

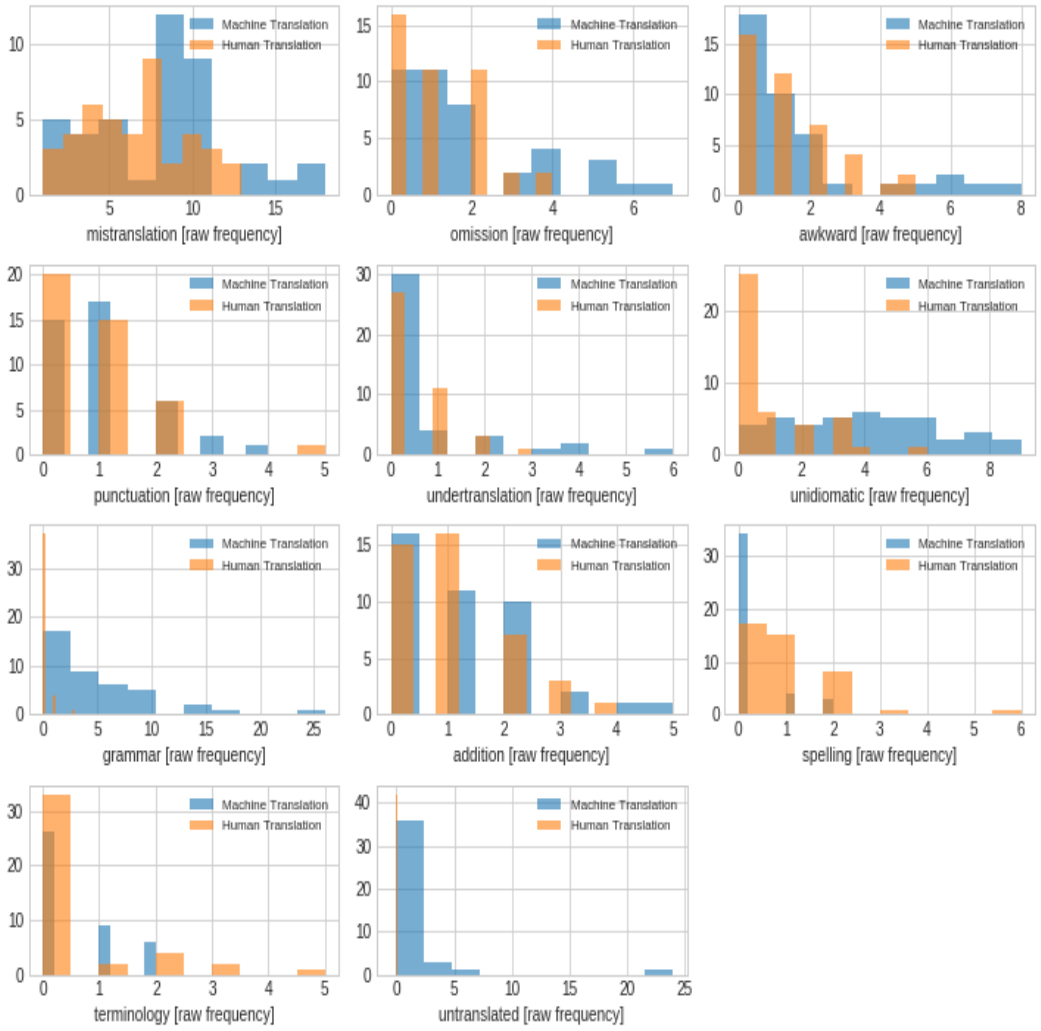


Figure 5.3 Translation Errors in HTs and MTs

5.4.1 PCA Analysis

As is mentioned earlier, I want to find out what are the major types of errors characteristic of either machine translations or human translations. For this purpose, I carried out PCA analysis with FactoMineR (Lê et al., 2008), using the counts of all error types as individual elements, the text types of target translation as

supplementary categorical variables[20] and the quality scores of each translation as supplementary quantitative variables. Considering the relatively small size of error types, I extract three principal components using the default rotation method[21], Table 5.16 presents the factor loadings of each error types on various dimensions (components), with high positive or negative loadings indicating the correlation between the corresponding error type and the component identified. On its basis, I name the first underlying dimension **'language misuse'**, and the second underlying dimension **'content inadequacy'**, and the third **'lexical mistakes'**. I also plot out the top 8 errors to the first two dimensions (two principal components) as shown in Figure 5.4.

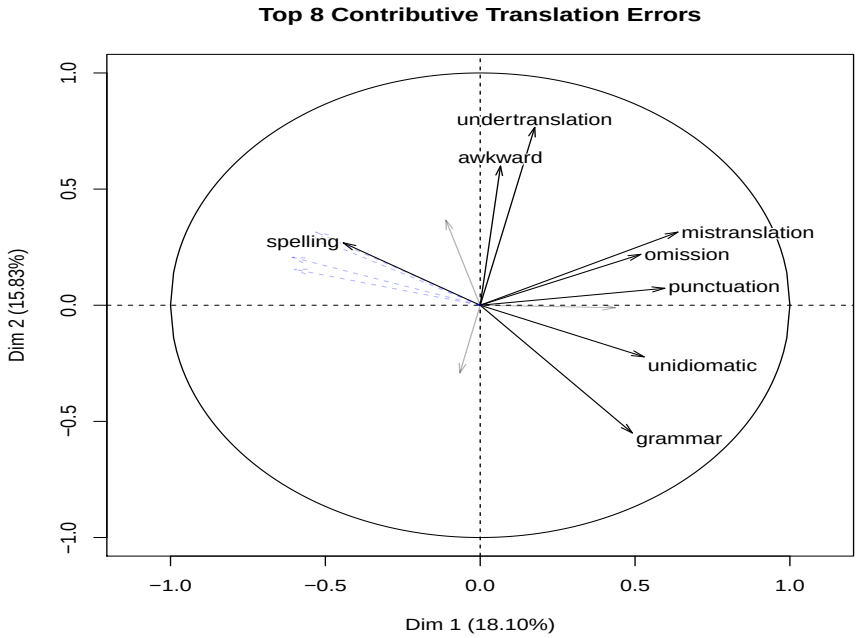


Figure 5.4 Top 8 Error Types in the First Two Dimension

Figure 5.4 basically conforms to the analysis at the beginning of this section: mistranslation, grammar and awkward are very differentiating features. In the first dimension, mistranslation, punctuation, grammar are positively prominent than other variables, and it is the same case with undertranslation and awkward in the second dimension. I conjecture that undertranslation and awkward are more about the translation strategy and competence (i.e. being unable to translate with ideal specificity and express the source content in a natural way), and mistranslation, punctuation and grammar are about the lexical, semantic and syntactic accuracy of the language in translations.

[20] A supplementary variable is a variable which will not be taken into account during the construction of the factorial axes, i.e. the calculation of distances between the individuals.

[21] i.e. varimax, an orthogonal method to scale the respective eigenvalues by the squared roots so as to obtain the eigenvectors as loadings.

	Dimension 1	Contribution	Cos2	Dimension 2	Contribution	Cos2	Dimension 3	Contribution	Cos2
mistranslation	0.638	20.462	0.407	0.315	5.692	0.099	-0.437	14.889	0.191
omission	0.519	13.519	0.269	0.219	2.759	0.048	0.381	11.346	0.145
awkward	0.066	0.216	0.004	0.599	20.622	0.359	0.150	1.763	0.023
punctuation	0.596	17.842	0.355	0.073	0.307	0.005	-0.050	0.194	0.002
undertranslation	0.176	1.559	0.031	0.766	33.700	0.587	0.066	0.342	0.004
unidiomatic	0.529	14.081	0.280	-0.222	2.836	0.049	0.333	8.645	0.111
grammar	0.491	12.115	0.241	-0.550	17.362	0.302	-0.001	0.000	0.000
addition	0.436	9.561	0.190	-0.010	0.006	0.000	-0.393	12.066	0.155
spelling	-0.442	9.811	0.195	0.269	4.147	0.072	-0.397	12.290	0.157
terminology	-0.111	0.617	0.012	0.366	7.700	0.134	0.574	25.715	0.329

Table 5.16 Factor Loadings of Errors Types on Different Dimensions

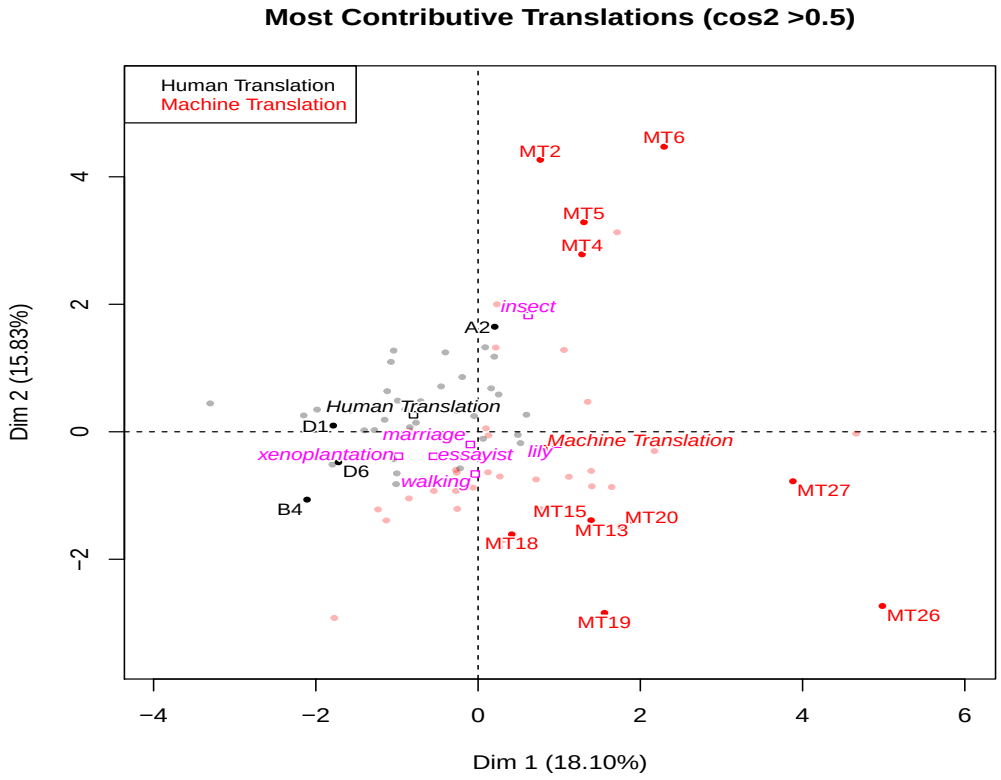


Figure 5.5 Distribution of Translations in the First Two Dimensions

To find out which dimension is more characteristic of HT or MT, I plot out the most contributive translations ($\cos 2 > 0.5$)[22]. Figure 5.5 indicates the interaction between the underlying components, individual translations and their types (domain, human or machine translation).

[22]The squared cosine shows how vital a dimension is to an individual observation. In our case, how likely a dimension characterizes a translation.

As is manifested in the figure, the first dimension, i.e. *language misuse*, characterizes most MTs, as top contributive translations to this dimension comprise mainly MTs (dark and light red dots). In contrast, HTs (black and grey dots) centre towards the second dimension, i.e. *content inadequacy*. These findings suggest that deficiency of HTs in quality may have to do with translators' inability of delivering the ST content in a sufficient manner. For MTs, these findings imply that language problems, such as grammaticality, naturalness, are typical. These findings are echoed by Vilar et al. (2006), who also maintain that language issues, such as wrong lexical choice, incorrect form, extra words, style and idiom, are the primary sources of Chinese-English errors.

Another finding from the graph is that translations of the source text of 'lily' (short story on perseverance) highly positively correlate with the first dimension, and in the similar vein, translations of the source text of 'insect' (science fiction) highly positively correlate with the second dimension, while translations of other topics/domains are more or less centring around the intersection of two dimensions with negative correlations. This finding suggests that heterogeneity of translation tasks may play a part in causing human translators and MT systems to produce errors of specific types. It also reminds us that during the feature engineering process specifications of the translation tasks should be taken into consideration, if possible. For instance, for science fiction, more terminology issues may be involved, but for casual texts on social life, such as essays, grammar and unidiomatic issues may be more prominent.

Now I move on to look into the interaction between the engineered features (discussed in Chapter 3) and quality scores (in Section 5.3.2), namely **UT**, **TS**, **IW**, **TM**, **AD**, **FL** and **TO**.

5.4.2 Weighting of Features to Translation Quality

In the WMT12 quality estimation shared task, the numbers of features range from dozens to several hundred. Some work has been devoted to the selection and integration of features to build predictive models (Luong et al., 2013; Shah et al., 2013; Stanojević and Sima'an, 2014). For HTQE, only a handful of similar work on feature engineering and selection is identified (Yuan et al., 2016), and there is a lack of in-depth analyses on the contribution of individual features to quality scores alongside the modelling process. Thus, I run a correlation analysis of the feature set described in Chapter 3. I use this criterion to rank the importance of individual features to the different components of translation quality discussed previously.

5.4.2.1 Number of Features

As listed in Section 3.2 of Chapter 3, I proposed a total number of 360 features for the document-level QE and 341 features for the sentence-level. The difference in

the number of the feature sets is due to that some features, such as the whole set of inter-sentential coherence and number of sentences in the document, do not apply at the sentence level. Unique document-level features are recorded in Appendix D.

5.4.2.2 Correlation Threshold

The Pearson correlation coefficient measures the strength of linear association between two variables (Sedgwick et al., 2012), in our case of HTQE, the association between any feature and each quality score (subcategorical or total).

As a rule of thumb, a conservative linear relationship exists if $|\mathrm{r}_{\mathrm{xy}} \geqslant \frac{2}{\sqrt{\mathrm{n}}}|$ (Krehbiel, 2004). Thus, for absolute r and the level of association, r below 0.20 indicates almost no linear relationship, if the sampled population is not very big (< 25). I select features which have a $\mathrm{r} > 0.30$ with each type of quality scores. In the following, I report the number of features which have an over-weak association with translation quality.

5.4.2.3 Correlation with Quality Scores

In the following, I report how a collection of features correlate with each type of quality components in term of the Pearson correlation coefficient. Features with $\mathrm{r} \geqslant 0.30$ are viewed as more contributive to each quality category in comparison to other features. Table 5.17 presents the specific number of different feature types selected for the 7 quality scores specified above. Note that for presentation and the convenience of analysis, I regroup the selected features into 11 subgroups. Most of these groups are self-explanatory. The group of bilingual distance features includes all log ratios of paired monolingual features and CBD measures. For a complete list of specific features selected for each type of quality components, refer to Appendix C.

It can be observed that features under the subcategories of alignment, log ratios and bilingual distance, constituency, dependency relations, pseudo-reference and back translation, semantic roles and shallow features are proportionally more salient for almost all quality types, given the overall number of features in those subcategory groups are limited (See section 3.2).

It is also found that cohesion and coherence features seem to be important for both 'content' and 'style' as more of such features are selected for *terminology*, *adequacy* and *fluency* that focus on both 'style' and 'content', while fewer such features are selected for *target mechanics* concerning language conventions.

Pseudo-reference and back-translation features are more associated with content adequacy than with language fluency as shown by the smaller number of them for *terminology*, *idiomatic writing*, *target mechanics* and *fluency*. This phenomenon could be possibly explained by the fact that these features are mainly MT metric scores to compare the lexical similarity and similarity measures based on continuous vectors measuring the semantic and syntactic relatedness.

Features class	UT	TS	IW	TM	AD	FL	TO
alignment	4	4	4	3	4	4	4
bilingual distance	51	51	45	39	52	39	52
cohesion and coherence	3	6	4	1	4	6	3
constituency	6	6	5	3	6	4	6
dependency relations	19	17	14	7	20	13	17
language modelling	1	1	2	1	1	1	1
POS Tags	7	6	4	2	6	5	5
Pseudo-reference and Back-translation	27	30	14	2	28	12	27
semantic role labels	4	3	4	1	3	3	3
shallow	5	7	5	4	6	5	6
Total	127	131	100	63	130	92	124

Table 5.17 Selected Features for Different Quality Scores ($|\mathrm{r}| \geqslant 0.3$)

Cohesion and coherence features capture the intra-sentential or inter-sentential linkage of segments (e.g. words, phrases, sentences, paragraphs). As a whole, these features should manifest to what extent how well the issues of formality, readability and naturalness are handled in the target translations. In a general case, well-connected and coherent sentences of translations read smooth and natural. In comparison, as regards the issue of semantic equivalence that is adequacy-related, features of bilingual distance and semantic roles show a higher affinity to *usefulness* and **adequacy**, two quality aspects indicating content completeness and accuracy, than to other quality scores as evidenced by the more substantial number of features under these categories.

It is also interesting to note that constituencies and dependency relations do not seem to favour either content or language partially. Instead, they demonstrate rather even distribution under the **UT**, **TS**, **IW**, **AD**, **FL** and **TO** scores. A possible explanation for this is that constituency features and dependency relations can capture correspondence between longer units of STs and TTs that carries semantic and syntactic information at the local level and incrementally across the document as a whole.

For illustration, I group and report features whose absolute correlations are above 0.5 following the framework proposed in Section 3.2. Features that correlate strongly with *usefulness*, *terminology*, *idiomatic writing*, *adequacy*, *fluency* and *total* are presented in Tables 5.12 to 5.17. As *target mechanics* only correlates strongly with one feature, i.e. target source object log ratio, it is not included.

Table 5.12 lists all the selected features in different categories that have demonstrated strong correlation with *usefulness*. It can be seen that length and complexity related shallow features (type, tokens and type-token ratio), word alignment and main content words and phrases (e.g. noun, noun phrase, verb phrase) have a strong positive correlation with *usefulness*, and bilingual distance features, mainly composed of log ratios comparing the frequencies, lengths, and statistics of corresponding features in STs and TTs, show strong negative correlation. Their correlation may be explained by the fact such words, phrases, and metrics are good indicators of

Feature class	Features	$\mathbf{r}$	$\mathbf{P}$ value
alignment	word alignment normalised by target length	0.60	
	word alignment normalised by source length	0.53	
bilingual distance	target source noun phrase log ratio	-0.52	
	target source empty words log ratio	-0.53	
	target source content words log ratio	-0.53	
	target source SRL A0 log ratio	-0.54	
	target source SRL A1 log ratio	-0.55	
	target language model log probability	-0.56	
	target source nominal modifier log ratio	-0.56	
	target source adverbial modifier log ratio	-0.57	
	target source verb phrase log ratio	-0.57	
	target source nominal subject log ratio	-0.59	
	source target type token ratio	-0.59	
	source target token log ratio	-0.60	
	source target type log ratio	-0.60	
	target source punctuation dependency log ratio	-0.60	
	target source conjunct log ratio	-0.60	
	source target LM probability log ratio	-0.62	
	target source object log ratio	-0.62	<0.0001
constituency	target verb phrase	0.55	
	target noun phrase	0.54	
dependency	target object	0.51	
	target punctuation dependency	0.50	
	target nominal modifier	0.50	
POS Tags	target nouns	0.56	
	target verbs	0.52	
	target punctuation	0.51	
semantic roles	target semantic roles (others)	0.54	
	target semantic roles (A1)	0.51	
shallow	target type token ratio	0.62	
	target tokens	0.57	
	target types	0.55	
	target content words	0.54	

Table 5.12 Contributive Features to Usefulness ($|\mathrm{r}| \geqslant 0.5$)

Feature class	**Features**	$\mathbf{r}$	$\mathbf{P}$ **value**
alignment	word alignment normalised by target length	0.58	
	word alignment normalised by source length	0.51	
bilingual distance	target source adverbial phrase log ratio	-0.51	<0.0001
	target source case log ratio	-0.51	
	target source linkings log ratio	-0.52	
	target source root log ratio	-0.53	
	target source empty word log ratio	-0.55	
	target source adjective modifier log ratio	-0.55	
	source target discourse CBD	-0.55	
	target source maker log ratio	-0.56	
	target source content words log ratio	-0.57	
	target source A1 log ratio	-0.57	
	target source A0 log ratio	-0.58	
	target source punctuation dependency log ratio	-0.60	
	target source nominal modifier log ratio	-0.60	
	target source noun phrase log ratio	-0.60	
	target source conjunct log ratio	-0.62	
	target source nominal subject modifier log ratio	-0.62	
	target source adverbial modifier log ratio	-0.62	
	source target types log ratio	-0.64	
	source target type toke ratio log ratio	-0.64	
	source target tokens log ratio	-0.64	
	target source verb phrase log ratio	-0.65	
	target source object log ratio	-0.66	
	source target LM probability log ratio	-0.67	
constituency	target noun phrase	0.52	
	target verb phrase	0.50	
language model	target language model log probability	-0.54	
POS tags	target nouns	0.51	
Pseudo-reference and back translation	Yandex back translation NIST score	0.53	
shallow	target type token ratio	0.61	
	target tokens	0.53	
	target types	0.51	

Table 5.13 Contributive Features to Terminilogy ($|\mathrm{r}| \geqslant 0.5$)

Feature class	Features	r	P value
bilingual distance	target source noun phrase log ratio	-0.50	<0.0001
	target source punctuation dependency log ratio	-0.51	
	target source conjunct log ratio	-0.52	
	source target types log ratio	-0.53	
	source target token log ratio	-0.54	
	source target type token log ratio	-0.54	
	target source adverbial modifier log ratio	-0.55	
	target source verb phrase log ratio	-0.56	
	target source object log ratio	-0.57	
	source target LM probability log ratio	-0.60	

Table 5.14 Contributive Features to Idiomatic Writing ($|\mathrm{r}| \geqslant 0.5$)

when transferred from the ST into the TT. Their distribution in an ST indicates the complexity and meaningfulness of the ST sentence, and their presence in the TT implies the degree of correspondence in both meaning and form, which are essential for the success of translation act.

Tables 5.12 to 5.17 show the strong correlative features with other quality aspects. We can observe that

- The majority of most contributive variables (features) come from the target side, and bilingual distance features. This finding implies that QE may be more dependent on the characteristics of target texts, and features that elaborate the ST and TT correspondence relation. However, the possibility that it might have been an artefact of the much smaller number of STs cannot be ruled out.
- Most features, regardless of POS, constituency and dependency relations, are content-related, e.g. noun, verb phrase, nominal modifier.
- bilingual distance features in the form of CBDs and log ratios of corresponding features show signs of strong negative correlation with the relevant quality aspect.
- some features contribute consistently to particular quality aspects. For all subcategorical scores, shallow features (e.g. target type-token ratio, target tokens, target types), dependency relations, constituencies (e.g. target noun, target verb) and alignment (e.g. one word alignment and two word alignment) have shown comparatively higher correlations with content-oriented translation quality: *usefulness*, *terminology*, *adequacy* and *total*. This consistency can be explained by the fact that these features are able to capture the length, complexity, syntax and semantics of both source texts and target texts.
- Among the features in each feature class, TT-side features demonstrate a higher correlation with quality scores than ST-side features. For example,

Feature class	**Features**	$\mathbf{r}$	$\mathbf{P}$ **value**
alignment	word alignment normalised by target length	0.62	
	word alignment normalised by source length	0.54	
bilingual distance	target source linkings log ratio	-0.52	
	target source root log ratio	-0.53	
	target source adjectival modifier log ratio	-0.53	
	target source marker log ratio	-0.53	
	target source empty words log ratio	-0.55	
	target source content words log ratio	-0.56	
	target source noun phrase log ratio	-0.57	
	target source SRL A1 log ratio	-0.57	
	target source SRL A0 log ratio	-0.57	
	target source nominal modifier log ratio	-0.60	
	target source adverbial modifier log ratio	-0.61	
	target source nominal subject log ratio	-0.62	
	target source punctuation dependency log ratio	-0.62	
	target source verb phase log ratio	-0.62	
	source target TTR log ratio	-0.63	<0.0001
	target source conjunct log ratio	-0.63	
	sourc target tokens log ratio	-0.63	
	source target types log ratio	-0.63	
	target source object log ratio	-0.65	
	source target LM probability log ratio	-0.66	
constituency	target noun phrase	0.55	
	target verb phrase	0.55	
	target adverbial phrase	0.51	
dependency	target nominal modifier	0.52	
	target object	0.52	
	target source case log ratio	-0.51	
language model	target LM probability log ratio	-0.57	
POS tags	target nouns	0.56	
	target verbs	0.52	
	target punctuation	0.51	
Pseudo reference and Back translation	Yandex corpus level NIST score	0.52	
semantic roles	target SRL others	0.54	
	target SRL A1	0.51	
shallow	target type toke ratio	0.64	
	target tokens	0.57	
	target types	0.55	
	target content	0.54	

Table 5.15 Contributive Features to Adequacy ($|\mathrm{r}| \geqslant 0.5$)

Feature class	Features	$\mathbf{r}$	$\mathbf{P}$ value
bilingual distance	target source punctuation dependency log ratio	-0.50	<0.0001
	target source nominal modifier log ratio	-0.50	
	target source nominal subject log ratio	-0.51	
	target source noun phrase log ratio	-0.52	
	target source conjunct log ratio	-0.52	
	source target types log ratio	-0.53	
	source target tokens log ratio	-0.54	
	target source adverbial modifier log ratio	-0.54	
	source target TTR log ratio	-0.54	
	target source verb phrase log ratio	-0.56	
	target source object log ratio	-0.57	
	source target LM probability log ratio	-0.58	

Table 5.16 Contributive Features to Fluency ($|\mathrm{r}| \geqslant 0.5$)

when it comes to the correlation with *Usefulness*, the selected TT-side features of constituencies, POS tags and semantic roles have the highest correlation coefficients, which suggests that the ST-side features do not have such high correlations. This phenomenon is applicable to all component scores and related to the process of human annotation where annotators tend to pay more attention to the quality of translations themselves. In other words, human raters may score translations without reliance on the STs. This hypothesis is confirmed by the interview with both annotators, who recall that in the beginning they read and compare the ST and TT back and forth, and gradually they start grading without using the ST once they think they are familiar with the ST content. They also admit that they assume the quality of ST is perfect and did not attempt to evaluate the STs with the same criterion. However, this explanation may warrant a more comprehensive investigation, such as a large-scale survey.

- Sentence-level correlation analysis shows a significant drop in Pearson correlation coefficient. I also carried out a sentence-level correlation analysis of the features with quality scores at the sentence level. However, we found no features showing correlation higher than the threshold $|\mathrm{r}| >= 0.3$. When it is set to be 0.2, there are only very few features weakly correlated with *terminology*, *target mechanics*, *adequacy* and *total*. It means that a majority of features have very weak correlation with the set of quality scores at the sentence level. I assume this may be caused by the data sparsity within the sentence-level translational data.

Feature class	**Features**	$\mathbf{r}$	**Pvalue**
alignment	word alignment normalised by target length	0.58	
	word alignment normalised by source length	0.51	
bilingual distance	target source linkings log ratio	-0.51	
	target source root log ratio	-0.53	
	target source adjectival modifier log ratio	-0.53	
	target source maker log ratio	-0.53	
	target source empty word log ratio	-0.54	
	target source SRL A1 log ratio	-0.55	
	target source content log ratio	-0.56	
	target source SRL A0 log ratio	-0.56	
	target source noun phrase log ratio	-0.57	
	target source nominal modifier log ratio	-0.59	
	target source punctuation log ratio	-0.60	
	target source nominal subject log ratio	-0.60	
	target source adverbial modifier log ratio	-0.61	<0.0001
	target source conjunct log ratio	-0.62	
	source target TTR log ratio	-0.62	
	source target type log ratio	-0.62	
	target source verb phrase log ratio	-0.62	
	source target tokens log ratio	-0.62	
	target source object log ratio	-0.65	
	source target LM probability log ratio	-0.66	
constituency	target noun phrase	0.52	
	target verb phrase	0.52	
language model	target LM probability	-0.54	
POS tags	target noun	0.52	
semantic roles	target SRL others	0.51	
shallow	target type token ratio	0.60	
	target tokens	0.53	
	target types	0.52	
	target content	0.51	

Table 5.17 Contributive Features to Total Score ($|\mathrm{r}| \geqslant 0.5$)

5.5 Summary

This chapter serves as an overarching introduction to our dataset, annotation scheme and data description. It is the first time that principal component analysis has been employed to study the distribution patterns of translation errors across MTs and HTs, together with their interaction with text types and topics. The pattern of HT errors (content inadequacy) implies that HT quality issues arise mainly due to either translators' decision-making (e.g. undertranslation is more a result of translation strategy) or their incapability of switching between two languages (e.g. awkward translations). In contrast, MT errors are more about language misuse. Natural language makes an obstacle to existing MT systems. Human translators, when translating into their native language, have fewer problems with it. This fact further suggests that the subtle difference between 'good' and 'bad' for human translations may be harder to determine. Translations of certain ST types may be more prominent on the underlying dimensions of errors identified via the PCA analysis. Such hidden dimensions are characteristic of either human translation or machine translation.

The contribution of the designed features to quality scores is also carefully examined via Pearson correlation. A rather comprehensive pairwise correlation analysis of all quality indicators with a total of 7 quality scores has been carried out. The most remarkable finding from the correlation analysis is that some categories of features contribute consistently more salient to all quality scores.

In next chapters, I will apply the whole set of engineered features to estimating human translation scores at the document and sentence level.

Chapter 6

Feature-based Document Level HTQE

6.1 Introduction

In this chapter, I focus on assessing human translations at the document level.

HTQE task is more challenging than MTQE as quality standards of individuals vary (Koponen et al., 2012; Turchi et al., 2013) and specifications of translation jobs change with domains, source text difficulty and target text quality. Apart from variation, judging the quality of human translations into one's native language is more difficult than for MTs. Automatically evaluating HTs through machine learning methods has been only recently proposed (Yuan et al., 2016).

The main contributions of this exploration are fourfold:

- I implement the reference-free quality estimation for human translation data at the document level. Different from MTQE, this is a task in the new domain.
- A broad set of features have been designed to capture the cross-sentence relations and contextual information which are less explored by the current MTQE research.
- This investigation is the largest scale fine-grained document-level HTQE to my knowledge. The learning goals of the proposed method are a collection of component quality scores on an extensive collection of translated documents by trainee translators. Fine-grained quality of individual documents is measured by direct assessment, which is believed to be highly reliable and cost-efficient (Graham et al., 2017b). Different from the effort scores (Callison-Burch et al., 2012), HTER scores and post-editing time (Bojar et al., 2013), the chosen scheme of quality annotations can provide detailed quality feedback closer to human evaluators, particularly for trainee translation evaluation.

- I investigate the plausibility of applying the proposed feature set to document-level MTQE by examining the performance of HTQE models on a collection of 42 human annotated MT-translated documents.

6.2 Related Work

The goal of this work is to predict the quality of human-translated documents at different granularities. That is to say, I estimate a hierarchy of quality scores from a weighted sum to subscores and subsubscores as defined in Chapter 5.3.2.

A considerable amount of work focus on quality prediction or confidence estimation at the word- and/or sentence-level, as part of a more extensive working pipeline, to filter MT translations for immediate use or post-editing. It was not until recently that TQE at the document level has become the focus of two DiscoMT workshop (Webber et al., 2013, 2015) and one of the main tracks of quality estimation shared tasks (Bojar et al., 2015, 2016a). Closely related to my goal are only a handful of work by Soricut and Echihabi (2010), Scarton and Specia (2014a), Scarton et al. (2015a, 2016), and Graham et al. (2017b). In Section 2.3.3, I gave a detailed introduction to the features used in some work involving document level MTQE. There are limitations to the current document-level MTQE. The main constraint on them is the scale of the research itself. More specifically, these work are generally at a limited scale. For example, Wong and Kit's work (2012) is based on the MetricsMATR 2008 development set (Przybocki et al., 2009a), which consists of only 5 documents. Recent document-level quality estimation subtask in ACL 2016 First conference on Machine Translation (WMT 2016) (Bojar et al., 2016a) has used only 146 translated documents, as opposed to the previous practice of using the paragraph as a substitute to whole texts (Bojar et al., 2014, 2015). The generalizability of metrics and models trained with such small sizes or pseudo data is likely to be undermined.

Besides, quality labels (scores) used in such shared tasks may be another issue. Instead of using direct human assessments, the organisers use METEOR scores (Banerjee and Lavie, 2005) computed against references as quality measures for the document-level development data. As METEOR is a unigram-based method of comparing lexical similarity, we have good reasons to believe this quality measure differs drastically from human judgement on the documents when textual cohesion and coherence are ignored.

In the next section, I will present the experiments using the engineered features proposed in Chapter 3 to build fine-grained predictive systems.

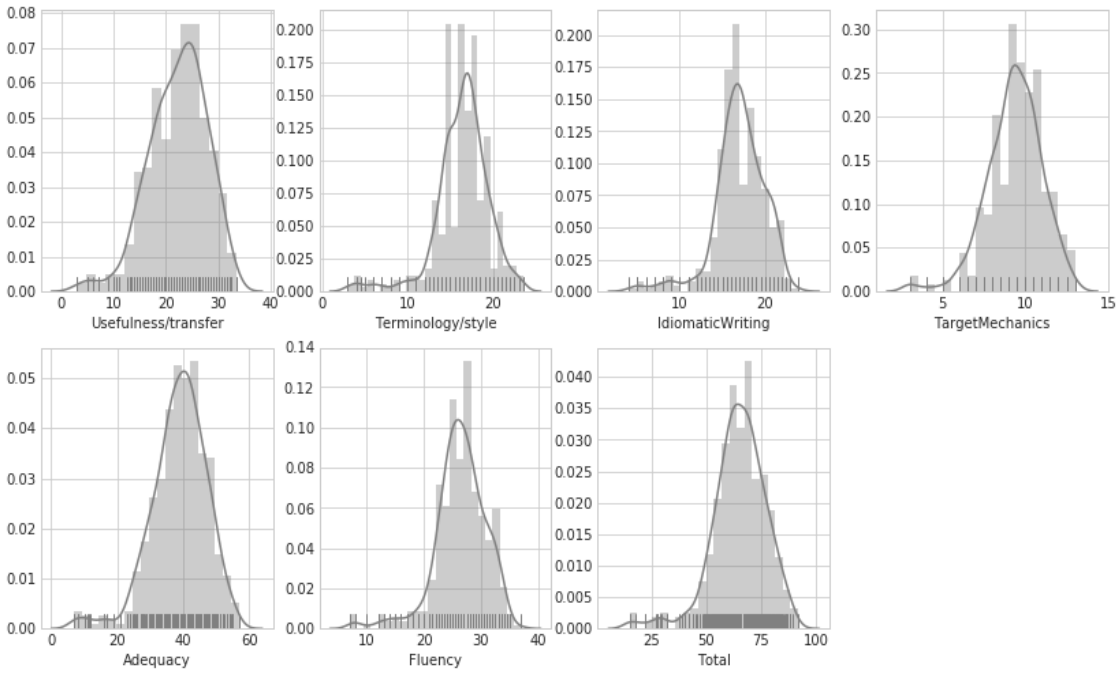


Figure 6.1 Distribution of Quality Scores (human annotations)

6.3 Experiment Setting

This section briefly describes the dataset, the machine learning algorithm and evaluation method I have chosen to build the quality prediction models.

6.3.1 Data

For the document-level quality estimation, I use the dataset described in Section 5.2. I split the data into two parts, 80% (365 documents) for training and 20% (92 documents) for testing.

As described in Section 5.2, the translated documents in the dataset are manually annotated with more than one quality score. A total score is composed of adequacy and fluency, two general criteria (Koehn and Monz, 2006), which are further split into four component scores assigned to each translated document per the ATA rubric. Thus, in this study, I try to predict 7 quality scores, i.e. **UT**, **TS**, **IW**, **TM** and the three derived scores **TO**, **AD**, **FL**.

Figure 6.1 indicates that quality labels for *Terminology* and *Fluency* scores have multiple peaks, in addition to the outliers of extremely low scores. Scores for *Usefulness* are rather left skewed, in contrast to scores for *Idiomatic Writing*. Such an imbalanced distribution and skewness may pose difficulties for the QE tasks.

6.3.2 Learning Algorithm:XGBoost

In this study building automatic quality estimation tasks has been cast as a regression learning task, in which translated documents are represented as a set of features that I have described in detail in Section 3.2 (illustrated by array 6.3.2),

$$\begin{bmatrix} \\ D_1 \\ D_2 \\ \cdots \\ D_{m-1} \\ D_m \end{bmatrix} = \overbrace{\begin{bmatrix} f_1 & \cdots & f_{n-1} & f_n \\ 0.2345 & \cdots & 0.2345 & 0.2345 \\ 0.2345 & \cdots & 0.2345 & 0.2345 \\ & \cdots & & \\ 0.2345 & \cdots & 0.2345 & 0.2345 \\ 0.2345 & \cdots & 0.2345 & 0.2345 \end{bmatrix}}^{features} \cdot \overbrace{\begin{bmatrix} S_1 & S_2 & \cdots & S_i \\ 30 & 20 & \cdots & 85 \\ 30 & 20 & \cdots & 85 \\ & & \cdots & \\ 30 & 20 & \cdots & 85 \\ 30 & 20 & \cdots & 85 \end{bmatrix}}^{\text{quality labels}}$$

where D_m denotes translated documents, f_n denotes each feature, and s_i denotes the number of quality scores to estimate. A chosen algorithm will then combine those features mathematically to form models capable of producing (a) final score(s) (s_i) for the input feature vectors.

For this study, I choose eXtreme Gradient Boosting (XGBoost) that implements the tree boosting learning method, which focuses on computational speed and model performance. XGBoost is a scalable end-to-end tree boosting system that has been used widely by data scientists to achieve the state-of-the-art results on Kaggle[1] competitions, e.g. AMS 2013-2014 Solar Energy Prediction Contest[2] (Chen and Guestrin, 2016). I choose this algorithm over others because it offers a combination of advanced features, which include models, e.g. regularised gradient boosting (Hastie et al., 2009), system parallelisation and sparse aware algorithms. The success of XGBoost is attributed to its innovative algorithmic optimisations. In particular, the novel gradient boosting tree learning algorithm helps train the model in an additive manner to handle data sparsity and solves the optimising difficulty. In the following, I use this algorithm to train the baseline and the QE models. For the experiments in this chapter, I implement regression on the scikit-learn interface of the XGBoost package.

However, the issue of sample size and dimensionality is problematic for the dataset used in this study. As described earlier, there are a total of 360 features for document representation and 341 for sentence representation. A large number of features pose a problem for short texts, e.g. incomplete translations and sentence translations, especially when there are only a small amount of training data. The relationship between the size of training data and the dimensionality has been studied extensively (Hughes, 1968; Kanal and Chandrasekaran, 1971; Fukunaga and Hayes, 1989). A general agreement is that the imbalance between the number of samples and the number of features, e.g. too many features for too few samples,

[1] https://www.kaggle.com/competitions
[2] https://www.kaggle.com/c/ams-2014-solar-energy-prediction-contest/

too many samples with too few features, is difficult for the induction, and there should be an optimum size of features. In case the number of features is larger than the 'optimal' size, feature selection (Yang and Pedersen, 1997; Miller, 2002) to choose a subset of features or sample selection (Mucciardi and Gose, 1971; Writh and Catlett, 1988) to choose appropriate training samples is advised. Also, the problem of data size and feature attributes is often handled by using cross-validation (Kohavi, 1995) that gives an insight on how a model generalises on an independent dataset. Alternatively, one can choose algorithms that are designed for high dimensionality. Support Vector Machine (SVM) (Cortes and Vapnik, 1995) and ensemble methods such as Random Forest (Ho, 1995) and XGBoost can have good performance for a larger number of features. In the domain of machine learning, as long as we make sure that the model works well through cross-validation, there is probably no need to worry about the number of features before pre-processing. I choose XGBoost partly for this reason.

In this study, I plan to predict the seven quality scores aforementioned, which are the continuous percentile scores. Therefore, I treat the task as a regression learning task.

6.3.3 Baseline

As HTQE is a rather new task, I use the proposed **QuEst** document-level basic features to build the baseline model. I implement the 17 top ranked features in the QuEst system as baseline features for a consecutive of quality estimation shared tasks (Callison-Burch et al., 2012; Bojar et al., 2016b, 2017). A brief introduction to the features has been given by Felice and Specia (2012). These features include:

- number of tokens in the source document
- number of tokens in the target document
- language model probability of source document
- language model probability of target document
- average source token length
- number of occurrences of the target word within the target hypothesis (averaged for all words in the hypothesis - type/token ratio)
- average number of translations per source word in the document (as given by IBM 1 table thresholded so that $\mathrm{prob}(\mathrm{t}|\mathrm{s}) > 0.2)$
- average number of translations per source word in the document (as given by IBM 1 table thresholded so that prob(t|s) > 0.01) weighted by the inverse frequency of each word in the source corpus

- percentage of unigrams in quartile 1 of the frequency (lower frequency words) in a corpus of the source language (SMT training corpus)
- percentage of unigrams in quartile 4 of the frequency (higher frequency words) in a corpus of the source document
- percentage of bigrams in quartile 1 of the frequency of source words in a corpus of the source language
- percentage of bigrams in quartile 4 of the frequency of source words in a corpus of the source language
- percentage of trigrams in quartile 1 of the frequency of source words in a corpus of the source language
- percentage of trigrams in quartile 4 of the frequency of source words in a corpus of the source language
- percentage of unigrams in the source sentence seen in a corpus (SMT training corpus)
- number of punctuation marks in the source sentence
- number of punctuation marks in the target sentence

QuEst baseline proves to be very strong as the systems built with these features rank higher than many other participating systems in the past quality estimation shared tasks (Samad Zadeh Kaljahi et al., 2013). For comparison, I extract the same feature set from the data and build a baseline model (hereinafter referred as **QuEst**) with these features.

6.3.4 Models

TopNcorr

Other than the **QuEst** basic features baseline, on the basis of the most important features selected by the correlation criterion ($\mathrm{r} \geqslant 0.3$) in Section 5.4.2, I train a model (**TopNCorr**) with these correlative features.

Full

Alongside the above two models, I also train a model (hereinafter referred as '**Full**') with all the proposed features (totalling 360) for the document-level HTQE in Chapter 3.

Kbest

In this study feature selection with univariate linear regression test is performed to select fewer features for model building. This feature selection method is done by computing the correlation between each regressor and the predicted target, and then convert the correlation to an F-score. In association with the grid search (Bergstra

and Bengio, 2012) that exhaustively generates candidate parameter values, this method is utilised to obtain the optimal number of features selected for each quality type. To this end, I empirically set the range of the optimal number of features to be $(20, 200)$ so that the maximum number of selected features does not exceed two-thirds of the total number of features. I eventually selected 73, 21, 20, 33, 31, 24, 23 features for **UT**, **TS**, **IW**, **TM**, **AD**, **FL**, **TO** respectively. Next, the selected top-k best features are fit in a model with fewer features. I name the trained model as **Kbest**.

For all of these four models, I apply a grid search with XGBboost in order to tune the hyper-parameters, setting a fixed learning rate of 0.05 and the number of trees 1000. The learning rate is a technique to slow down the learning in the gradient boosting model by applying a shrinkage factor set less than 1.0 so as to make fewer corrections for each tree added to the model. It is usually set in a range of 0.1-0.3 or less than 0.1. Most gradient boosting based methods are configured by default with a relatively small number of trees (e.g. hundreds or thousands). In this study, I set a fixed learning rate of 0.05 and the number of trees 1000, while tuning for other hyper-parameters of a specified range[3] via grid search.

6.3.5 Evaluation

Other than the *MSE* metric for model selection, the models are also presented with the correlation with human judgement and confidence intervals. *MSE* assesses the quality of predictor in the form of

$$\mathrm{MSE} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{Y_i} - \hat{\mathrm{Y}}_\mathrm{i})^2,$$

where $(\mathrm{Y_i} - \hat{\mathrm{Y}}_\mathrm{i})^2$ denotes the squares of errors between the true sample value and the predicted value, n represents the total number of samples.

However, for the cross-model comparison purpose, I prefer correlation as it does not need a conversion to be comparable for different response variables, e.g. **UT** and **TS** are on different scales. In this study, three most commonly used correlation coefficients are used, namely *Pearson correlation coefficient*, *Spearman correlation* and *Kendall's Tau*. *Pearson* r is used as the primary criterion for comparison.

6.4 Results and Discussion

6.4.1 Hyper-parameters

I list the hyper-parameters tuned for each model (including the baseline) through grid search in Table 6.1. These hyper-parameters are used for training the optimal

[3] For example, I specify the size of decision three as 3-6.

bf Model	**Hyperparameters**	**UT**	**TS**	**IW**	**TM**	**AD**	**FL**	**TO**
QuEst	colsample_bytree	0.5	0.5	0.5	0.5	0.8	0.5	0.6
	max_depth	3	3	3	3	3	3	3
	min_child_weight	6	5	3	3	6	3	3
	subsample	0.8	0.5	0.7	0.8	0.5	0.7	0.7
Full	colsample_bytree	0.5	0.7	0.7	0.7	0.7	0.5	0.6
	max_depth	6	5	5	5	6	6	4
	min_child_weight	3	5	6	4	4	4	4
	subsample	0.8	0.6	0.7	0.7	0.5	0.8	0.5
TopNCorr	colsample_bytree	0.7	0.6	0.5	0.5	0.6	0.5	0.6
	max_depth	5	6	6	6	5	6	3
	min_child_weight	6	4	6	3	5	6	6
	subsample	0.6	0.7	0.8	0.7	0.6	0.7	0.5
Kbest	colsample_bytree	0.7	0.8	0.6	0.7	0.7	0.5	0.5
	max_depth	4	5	6	5	6	5	6
	min_child_weight	4	3	6	6	4	4	3
	subsample	0.6	0.7	0.6	0.5	0.5	0.8	0.8

Table 6.1 Model Hyperparameters

models in comparison, which are then applied to predict the component quality scores aforementioned on the testing data (92 documents). Details of parameters for these models are given in Table 6.1. To cut off the running time, I carried out grid search for optimal parameters within a specified range only for the XGBoost model. I set a fixed learning rate $\eta = 0.01$ and tune the Booster parameters instead:

- $\mathbf{min\_child\_weight}$ which defines the minimum sum of weights of all observations in a child to control over-fitting. Higher values prevent the model from learning relations specific to the training samples and may lead to under-fitting. In this study, I set a range of $[5, 8]$ for grid search.
- $\mathbf{max\_depth}$ which defines the maximum depth of a tree, typically ranging between 3 and 10. Higher values will increase the model complexity, leading to overfitting. In this study, I set a range of $[5, 8]$ for grid search.
- $\mathbf{colsample\_bytree}$ which denotes the subsample fraction of columns to be randomly selected for each tree, ranging between $(0, 1]$. In this study, I set this range to be $[0.5, 0.8]$ for grid search.
- $\mathbf{subsample}$ which indicates the subsampling ratio of the training instance, usually ranging between 0.5 and 1. Too low values may lead to underfitting. In this study, I set the range to be $[0.5, 0.8]$ for grid search.

6.4.2 Best Features

As mentioned earlier, I carried out k-best feature selection in association with the grid search, and I found the optimal number features for each type of quality scores

respectively. These features are given per their importance to the **Kbest** model in the form of F statistics. Basically, an ANOVA statistical test is taken to measure the individual effect of each of many regressors (for regression) and returns the F-value between the label and feature that indicates the significance of the feature to the model. A simple criterion is that the higher the value of F statistics, the more contributive the feature is to the model. The selected features for each type of quality score in the model are given in Tables 6.2 to 6.8.

Feature class	Features	F-statistic
alignment	Word alignment normalized by target length	194.89
	word alignment normalized by source length	128.84
	two word alignment normalized by target length	121.28
	two word alignment normalized by source length	78.73
bilingual distance	source target LM log probability log ratio	203.8
	target source object dependency log ratio	198.38
	source target types log ratio	187.02
	source target tokens log ratio	186.57
	target source conjunct log ratio	181.22
	source target type toke ratio log ratio	181.02
	target source verb phrase log ratio	170.74
	target source adverbial modifier log ratio	170.61
	target source punctuation dependency log ratio	170.08
	target source nominal subject log ratio	157.9
	target source nominal modifier log ratio	140.56
	target source SRL A0 log ratio	140.45
	target source content words log ratio	131.54
	target source SRL A1 log ratio	129.82
	target source noun phrase log ratio	125.38
	target source empty words log ratio	116.94
	target source root log ratio	105
	target source linkings log ratio	99.62
	target source adverbial phrase log ratio	98.42
	target source adjectival modifier log ratio	94.98
	target source case log ratio	86.26
	target source markers log ratio	84.7
	target source SRL other log ratio	82.13
	source target CBD	78.19

Continued on next page

Table – *Continued from previous page*

Feature class	Features	F-statistic
	target source logical connectives log ratio	69.51
	target source prepositional phrase log ratio	63.09
	target source demonstrative log ratio	61.57
Cohesion and Coherence	target logical connectives	69.11
Constituency	target verb phrase	133.79
	target noun phrase	127.34
	target adverbial phrase	114.24
	target prepositional phrase	64.71
	target adjectival phrase	63.11
Dependency	target adverbial modifier	110.5
	target nominal modifier	105.8
	target punctuation dependency	102.08
	target object dependency	98.81
	target adjectival modifier	88.48
	target nominal subjects	87.99
	target case	87.16
	target root	85.69
	target phrasal verb particle	78.69
	target compound	75.67
	target conjunct	65.88
	Target marker	64.68
language model	target LM log probability	122.72
POS tags	target nouns	148.1
	target adverbs	109.11
	target verbs	105.82
	target punctuation marks	103.66
	target adjectives	75.91
pseudo-reference and Back translation	Yandex pseudo-reference corpus level NIST score	90.48
	Google back translation Levenshtein 2	76.96
	Bing pseudo-reference Levenshtein 2	75.72
	Yandex back translation Levenshtein 2	75.53
	Google pseudo-reference corpus level NIST score	66.84
	Yandex pseudo-reference RIBES score	65.12
	Google pseudo-reference Sorensen	64.97

Continued on next page

Table – *Continued from previous page*

Feature class	Features	F-statistic
	Google pseudo-reference Levenshtein 2	64.41
	Bing pseudo-reference Sorensen	62.12
semantic roles	target SRL others	133.38
	target SRL A1	101.9
	Target SRL A0	66.26
shallow	target TTR	178.4
	target tokens	140.02
	target content words	138.92
	target types	129.98
	source average sentence length	66.83
	target empty words	61.12

Table 6.2 Features for Usefulness in **Kbest**

Feature class	Features	F-statistic
alignment	word alignment normalized by target length	171.78
bilingual distance	source target LM log probability log ratio	251.86
	target source verb phrase log ratio	237.82
	target source object log ratio	222.71
	target source adverbial modifier log ratio	218.01
	source target tokens log ratio	213.84
	source target TTR log ratio	213.56
	source target types log ratio	208.21
	target source nominal subject log ratio	187.39
	target source conjunct log ratio	182.34
	target source noun phrase log ratio	179.88
	target source nominal modifier log ratio	166.6
	target type token ratio	166.11
	target source punctuation dependency log ratio	159.09
	target source SRL A0 log ratio	149.9
	target source content word log ratio	142.62
	source target shallow features CBD	140.29
	target source SRL A1 log ratio	130.67
	target source marker log ratio	119.11
	target source empty words log ratio	117.81
	target source adjectival modifier log ratio	117.04

Table 6.3 Features for Terminology in **Kbest**

Feature class	Features	F-statistic
alignment	word alignment normalized by target length	70.64
bilingual distance	source target LM log probability log ratio	143.24
	target source object log ratio	125.24
	target source adverbial modifier log ratio	120.86
	target source verb phrase log ratio	115.99
	source target TTR log ratio	99.62
	source target tokens log ratio	97.26
	target source conjunct log ratio	95.57
	source target types log ratio	92.91
	target source noun phrase log ratio	85.01
	target source demonstrative log ratio	84.26
	target source punctuation dependency log ratio	83.49
	target source nominal subject log ratio	82.57
	target source nominal modifier log ratio	79.03
	target source markers log ratio	73.98
	target source SRL A0 log ratio	70.31
	target source content log ratio	69.66
	target source adverbial phrase log ratio	66.36
	target source logical connective log ratio	65.87
shallow	target type token ratio	77.89

Table 6.4 Features for Idiomatic Writing in **Kbest**

Feature class	Features	F-statistic
alignment	word alignment normalized by target length	63.54
	word alignment normalized by source length	48.07
	two word alignment normalized by target length	39.49
bilingual distance	target source object log ratio	90.05
	target source noun phrase log ratio	87.15
	target source verb phrase log ratio	83.2
	source target LM log probability log ratio	80.3
	source target TTR log ratio	77.82
	source target tokens log ratio	76.2

Continued on next page

Table – *Continued from previous page*

Feature class	Features	F-statistic
	target source adverbial modifier log ratio	75.43
	source target types log ratio	73.08
	target source nominal subject log ratio	72.41
	target source SRL A0 log ratio	71.52
	target source nominal modifier log ratio	68
	target source conjunct log ratio	66.66
	target source content log ratio	66.55
	target source adjectival modifier log ratio	54.19
	target source adverbial phrase log ratio	51.99
	target source punctuation dependency log ratio	50.39
	target source empty word log ratio	49.07
	target source root log ratio	47.79
	source target shallow features CBD	44.13
	target source marker log ratio	43.91
	target source demonstrative connective log ratio	41.33
	target source SRL A1 log ratio	40.56
	target source average sentence length log ratio	38.88
	source target discourse CBD	37.93
cohesion and coherence	source adjacent sentence overlapping	46.96
constituency	target adverbial phrase	42.53
POS tags	target adverbials	38.43
shallow	source average sentence length	53.97
	target TTR	47.68
	target content words	38.62

Table 6.5 Features for Target Mechanics in **Kbest**

Feature class	Features	F-statistic
alignment	word alignment normalized by target length	206.79
	word alignment normalized by source length	136.33
	two word alignment normalized by target length	128.7
	source target LM log probability log ratio	245.77
	target source object log ratio	230.57
	source target tokens log ratio	217.99

Continued on next page

Table – *Continued from previous page*

Feature class	Features	F-statistic
bilingual distance	source target types log ratio	216.19
	target source verb phrase log ratio	213.77
	source target TTR log ratio	213.55
	target source adverbial modifier log ratio	206.89
	target source conjunct log ratio	201.48
	target source nominal subject log ratio	185.8
	target source punctuation dependency log ratio	183.63
	target source nominal modifier log ratio	164.73
	target source SRL A0 log ratio	158.31
	target source noun phrase log ratio	157.31
	target source content log ratio	148.86
	target source SRL A1 log ratio	142.88
	target source empty word log ratio	128.41
	target source root log ratio	117.31
constituency	target verb phrase	133.25
	target noun phrase	131.74
	target adverbial phrase	119.17
dependency	target nominal modifier	115.72
language model	target LM log probability	128.9
POS tags	target nouns	144.78
semantic roles	target SRL others	130.68
shallow	target TTR	192.59
	target tokens	141.89
	target content words	135.29
	target types	129.41

Table 6.6 Features for Adequacy in **Kbest**

Feature class	Features	F-statistic
alignment	word alignment normalized by target length	73.95
constituency	target adverbial phrase	58.84
	source target LM log probability log ratio	128.45
	target source object log ratio	122.22
	target source verb phrase log ratio	112.94
	target source adverbial modifier log ratio	112.19
	source target TTR log ratio	99.7
	source target tokens log ratio	97.4

Continued on next page

Table – *Continued from previous page*

Feature class	Features	F-statistic
bilingual distance	target source noun phrase log ratio	93.58
	source target types log ratio	93.11
	target source conjunct log ratio	91.87
	target source nominal subject log ratio	85.85
	target source nominal modifier log ratio	81.56
	target source SRL A0 log ratio	76.94
	target source punctuation dependency log ratio	76.31
	target source content log ratio	74.5
	target source demonstrative log ratio	72.03
	target source marker log ratio	67.19
	target source adverbial phrase log ratio	66.11
	target source adjectival modifier log ratio	61.72
	target source shallow features CBD	59.86
	target source root log ratio	58
	target source empty words log ratio	56.3
shallow	target type token ratio	71.49

Table 6.7 Features for Fleuncy in **Kbest**

Feature class	Features	F-statistic
alignment	word alignment normalized by target length	164.7
	word alignment normalized by source length	112.99
	two word alignment normalized by target length	108.43
bilingual distance	source target LM log probability log ratio	223.19
	target source object log ratio	210.03
	target source verb phrase log ratio	194.08
	target source adverbial modifier log ratio	189.38
	source target tokens log ratio	186.93
	source target TTR log ratio	185.81
	source target types log ratio	183.08
	target source conjunct log ratio	173.74
	target source nominal subject log ratio	160.72
	target source punctuation dependency log ratio	153.41
	target source noun phrase log ratio	147.63

Continued on next page

Table – *Continued from previous page*

Feature class	Features	F-statistic
	target source nominal modifier log ratio	145.52
	target source SRL A0 log ratio	138.92
	target source content words log ratio	131.77
	target source SRL A1 log ratio	115.04
	target source empty words log ratio	108.84
POS tags	target noun	111.02
shallow	target TTR	155.21
	target tokens	111.19
	target content words	107.01

Table 6.8 Features for Total Scores in **Kbest**

These selected features are grouped by the categories adopted in Section 5.4.2.3. As we can see, alignment features, shallow features and bilingual distance features consisting of paired monolingual feature log ratios and distance measures such as CBD are most common features for all quality aspects in the **Kbest** models. It is worth mentioning that alignment and bilingual distance features are selected for all quality types. Their selection as best quality predictive features can be explained by the fact that these two groups of features can capture the relationship of correspondence between the STs and TTs. Notably, specific features under each group (excluding bilingual distance) comprise overwhelmingly target side features. The latter finding is consistent with what I have found through the paired correlation analysis in Chapter 5. These two findings conform to our intuition that translation is a reproduction of the STs in the target language, and there exists a mapping of the major elements which are embodied lexically, semantically and syntactically. To measure the translation quality in some sense is to measure the degree of mapping. It generally holds true if we leave aside the element of creativity. On a different note, when different translations are compared, the distribution of representations in the TTs is more important to differentiate them regarding quality, since source-side representations remain stable for translations of the same TT.

Other findings include:

- for **UT**, It is observed that most subcategories of features are selected, and these features, including pseudo-reference and back translations, semantic roles and shallow features, are useful to capture the completeness of meaning transfer in different manners. Most notably, this group has the largest number of features selected, indicating the characteristic of its overall evaluation of completeness.

- for **TS**, the selected features suggest that bilingual distance and alignment features are useful to represent the lexical equivalence between STs and TTs.

- for **IW**, LM log probability log ratio of both ST and TT stands out among other features because it proves effective to measure language quality. Log ratios of their feature pairs include semantics of content words and phrases (e.g. noun phrase), cohesion and coherence devices, indexes of language complexity (e.g. TTR), which are essential to construct a meaningful and fluent utterance.
- for **TM**, other than those log ratios capturing correspondence, local or long-distance dependency relations such as constituency and cohesion and coherence devices, in addition to shallow features of length and complexity, are selected because they are concerned with language use. In particular, both the adverbial POS tags and constituencies are selected. I postulate that the use of adverbials in the translations helps capture the peculiarity of language use, which demonstrates the translator's target language competence (Pérez-Paredes and Sánchez-Tornel, 2014).
- Overall, selected features for different quality scores, including the four sub-component scores and the three derivative scores, share a lot of overlapping. This phenomenon suggests that features in my framework can capture multiple aspects of quality.

6.4.3 Model Performance

Model	# Features	Quality	MSE	r	ρ	τ
QuEst	17	UT	23.95	0.58	0.39	0.27
		TS	6.13	0.7	0.3	0.21
		IW	5.69	0.7	0.3	0.21
		TM	2.36	0.54	0.21	0.15
		AD	50.08	0.63	0.38	0.27
		FL	13.58	0.67	0.27	0.19
		TO	101.87	0.67	0.33	0.24
Full	328	UT	21.59	0.61	0.42	0.3
		TS	5.64	0.72	0.37	0.26
		IW	5.16	0.73	0.39	0.28
		TM	1.83	0.64	0.38	0.28
		AD	44.08	0.68	0.45	0.32
		FL	11.42	0.73	0.43	0.32
		TO	87.28	0.72	0.44	0.32
TopNcorr	127	UT	21.99	0.62	0.4	0.28
	131	TS	5.84	0.72	0.37	0.26
	100	IW	4.73	0.76	0.44	0.32
	63	TM	1.74	0.66	0.39	0.28
	130	AD	47.32	0.66	0.39	0.28
	92	FL	11.41	0.73	0.37	0.27
	124	TO	95.72	0.7	0.36	0.26
Kbest	73	UT	21.18	0.62	0.42	0.30
	21	TS	5.22	0.74	0.41	0.29
	20	IW	5.07	0.73	0.39	0.29
	33	TM	1.79	0.65	0.39	0.29
	31	AD	40.96	0.7	0.48	0.35
	24	FL	11.32	0.73	0.39	0.28
	23	TO	85.68	0.72	0.45	0.33

Table 6.9 Feature-based Document-level Quality Estimation Models

As shown in Table 6.9, **Full** model has shown significant improvement over the **QuEst** baseline with reduced MSEs and increased correlations with human annotations for 7 types of quality scores. **TopNCorr** model, although it has outperformed **QuEst** on all aspects of quality, only does better in estimating *idiomatic writing* and *target mechanics* than **Full** using the full feature set. When k-best feature selection is employed in combination with grid search, **Kbest** model using fewer features has achieved slightly better performance than **Full**. Therefore, we are confident to say that models built with the proposed features outperform the one built with QuEst feature set.

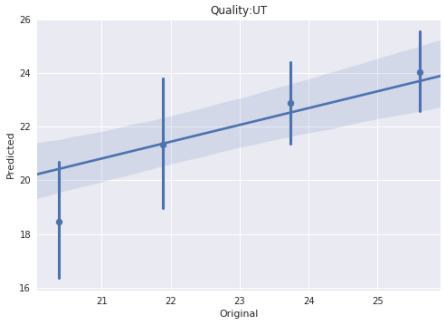
I further explored the central tendency of prediction by all four models, grouping all estimates into 4 bins to approximate the four quartile distribution of the testing data. The performance of each model for each type of quality score is visualised with

a regression plot, in which the line through the shadowed area (at confidence interval 95%) is a regression line representing the means of bootstrapping estimators, and four vertical lines are the binned original scores on the x-axis. The first bin to the left is made up of lowest quality scores, and the fourth bin to the right top quality scores in the same type (e.g. Usefulness or Adequacy). Similarly, the second bin and third bin in the middle represent the intermediate scores between the upper bound and lower bound of the first and the fourth bin. On the y-axis of the plot are the predicted scores by the trained model. For example, Figure 6.2 shows the central tendency of **QuEst** model on the same testing data (92 documents).

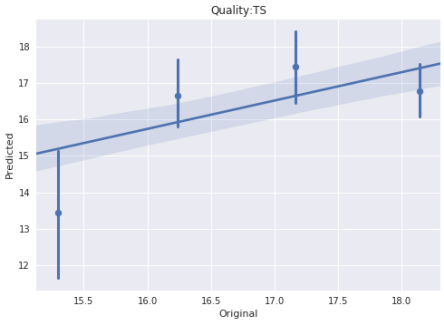
Because of the intrinsic flaws of assessing the tails of quality range, supervised systems have difficulty in assessing the very 'bad' and very 'good' translations (Moreau and Vogel, 2014). I will focus my discussion on the prediction of scores in the first bin and the fourth bin, as the performance of models in the first bin indicates their capability of estimating the quality for terrible translations, and their performance in the fourth bin indicates how well they predict the good translations.

Except for **Usefulness** (Figure 6.2a), **Idiomatic Writing** (Figure 6.2c) and **Adequacy** (Figure 6.2e), the **QuEst** model has shown the tendency to overestimate the corresponding scores in the fourth bin as the average (midpoint) of the binned scores falls far below the regression line, while consistently overestimating the lower scores in the first bin for all quality types.

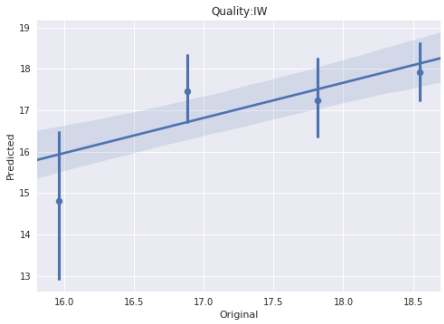
For the **Full** model, one of the noticeable patterns of its predictions is that the model consistently predicts well the fourth bin quality scores which are mostly overestimated by **QuEst** model. In the meantime, the **Full** model is more precise in predicting the second, third and fourth bin scores of the testing data (Figures 6.3c to 6.3g). However, a clear overestimate of the first bin scores is also observed. In general, the **Full** model has narrower corresponding confidence intervals in comparison to the **QuEst** model. This tendency further confirms its higher precision as demonstrated by its lower MSEs and the increased correlation.



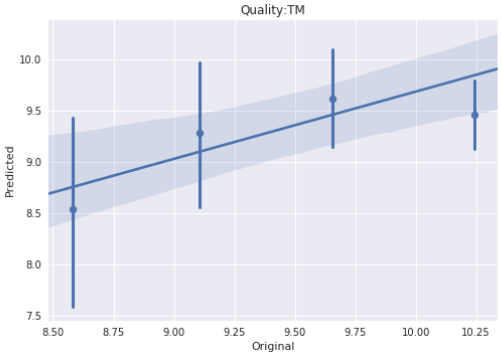
(a) Usefulness



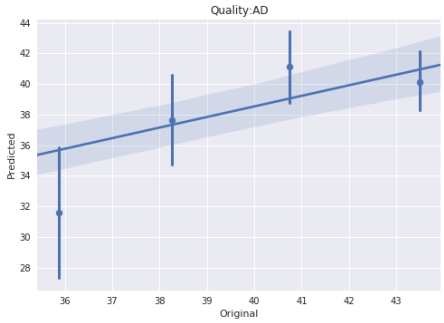
(b) Terminology



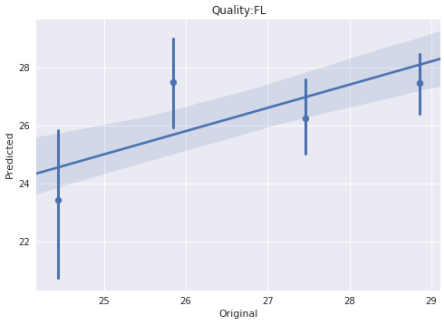
(c) Idiomatic Writing



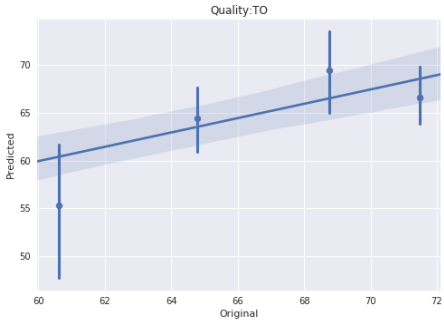
(d) Target Mechanics



(e) Adequacy

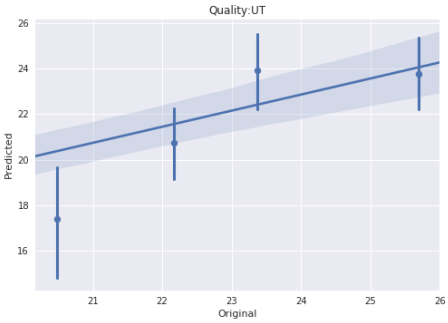


(f) Fluency

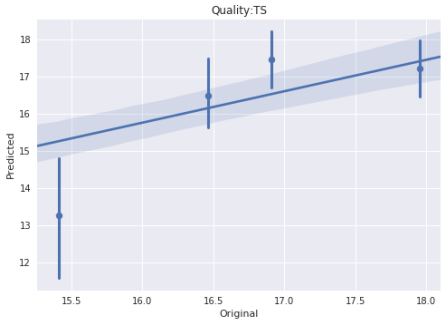


(g) Total

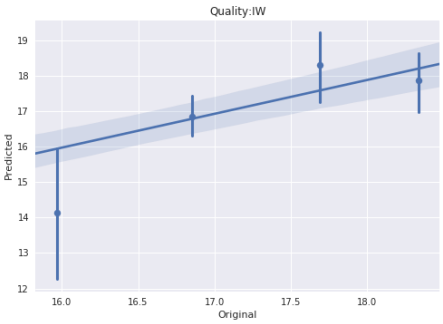
Figure 6.2 Regression Plots of QuEst Model on Testing Data



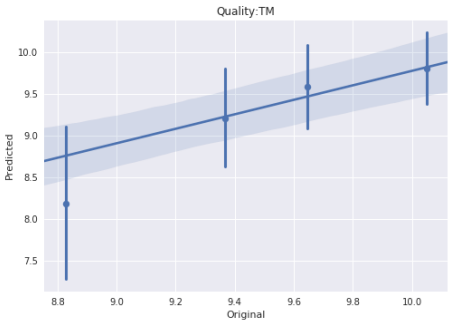
(a) Usefulness



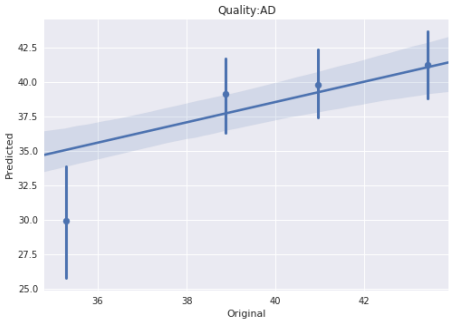
(b) Terminology



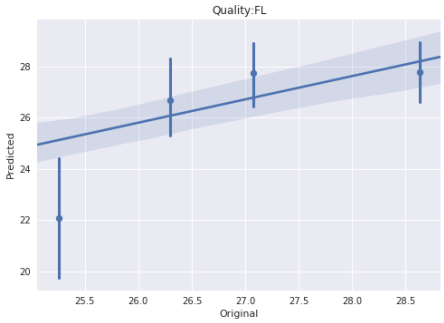
(c) Idiomatic Writing



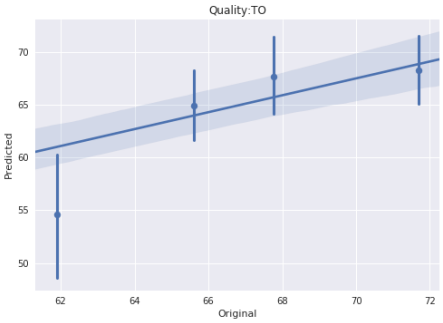
(d) Target Mechanics



(e) Adequacy

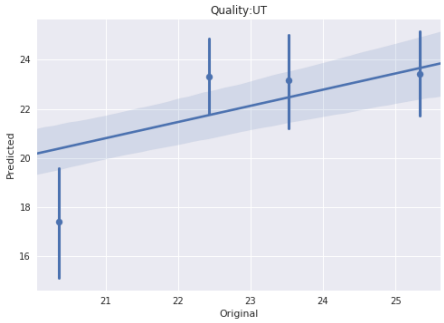


(f) Fluency

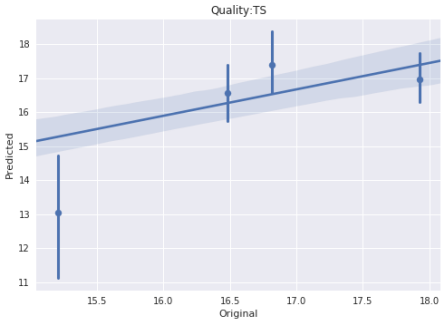


(g) Total

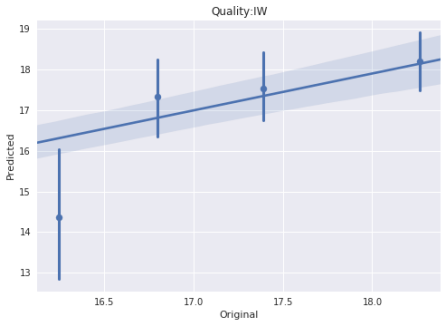
Figure 6.3 Regression Plots of Full Model on Testing Data



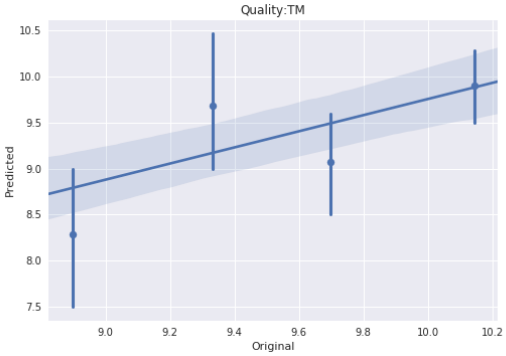
(a) Usefulness



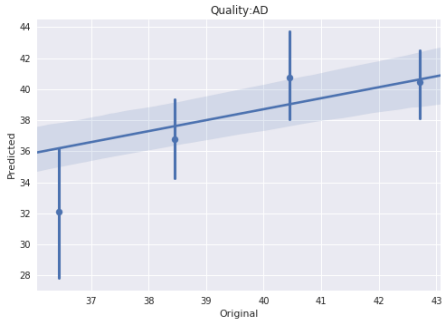
(b) Terminology



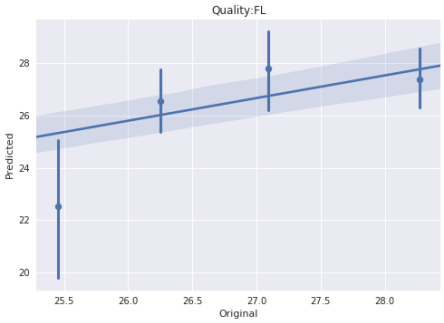
(c) Idiomatic Writing



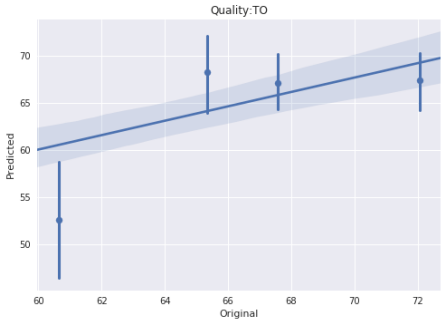
(d) Target Mechanics



(e) Adequacy

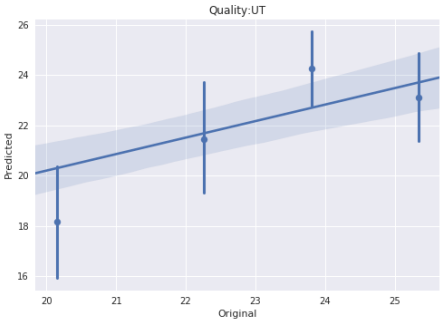


(f) Fluency

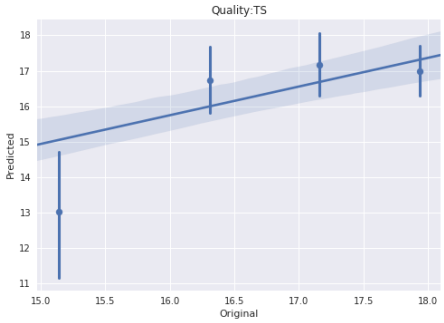


(g) Total

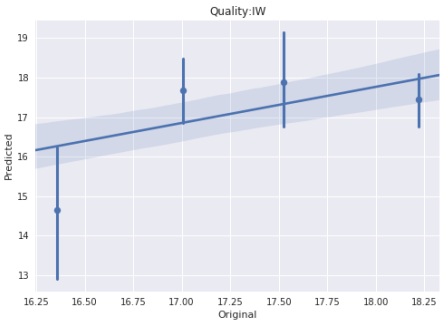
Figure 6.4 Regression Plots of TopNCorr Model on Testing Data



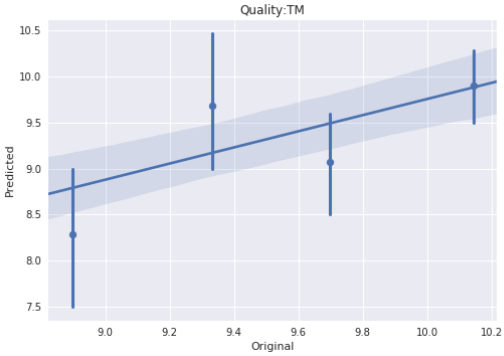
(a) Usefulness



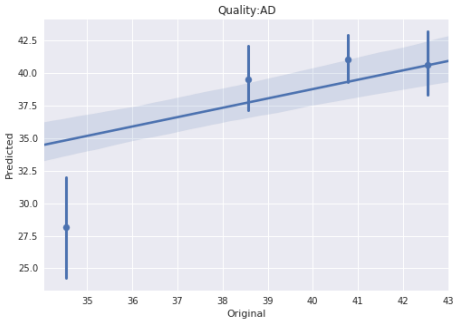
(b) Terminology



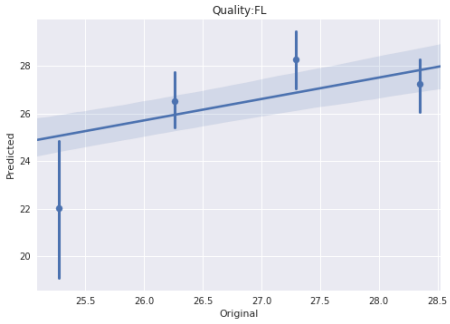
(c) Idiomatic Writing



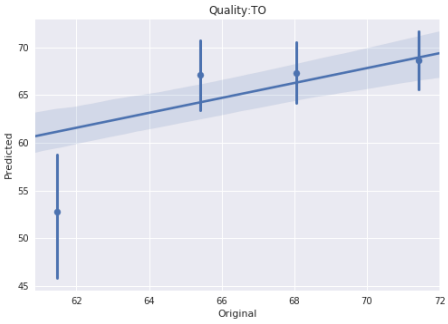
(d) Target Mechanics



(e) Adequacy



(f) Fluency



(g) Total

Figure 6.5 Regression Plots of Kbest Model on Testing Data

Feature	Freq. in Example 1	Freq. in Example 2	log-likelihood ratio		sig.
dependent	5.22	62.16	10.48		**
noun compound	3.48	68.08	17.19		***
coordination	9.57	100.64	14.72		***
case	0.87	38.48	13.82		***
adverbial phrase	3.44	33.48	4.43		*
adverbial	11.18	53.01	0.99		0.32
conjunct	1.72	39.06	10.75		**
Determiner (dependency)	1	5.58	0.22		0.64
noun	18.92	231.57	40.18		***
adjectival phrase	0.78	13.26	3.05		0.08
noun phrase	28.86	240.89	25.21		***
prepositional phrase	3.12	24.31	13.82		***
verb phrase	17.16	163.54	4.43		*
adjectives	0.86	16.74	0.99		0.32
SRL A0	6.88	47.43	10.75		**
SRL A1	8.6	75.33	0.22		0.64
SRL A2	1.72	13.95	40.18		***
SRL others	18.92	153.45	3.05		0.08
nominal modifier	8.7	94.72	0.40		0.53
nominal subject	7.83	82.88	25.21		***

Table 6.10 Feature Difference in Example 1 and 2.

Overall, **Kbest** continues the tendency of the better prediction for quality scores in the fourth bins. The fourth bin plot largely overlaps with the confidence interval for the upper bound scores under each quality category. It consistently overestimates the scores in the first bins as the first bin plot falls far below the confidence interval area for the lower bound scores.

In general, models built with the proposed features, including the **Full** model, **TopNCorr** model and the **Kbest**, have shown improved performance in estimating above-average quality scores (in the 2-4th bins). However, all of these models, including **QuEst**, are likely to elevate the scores of 'bad' translations for all the 7 quality types. I will illustrate this point with the estimations for two translations of the same ST (See Table A.1 in Appdendix A). Note that the highlighted red words imply the deficiency in **UT**, blue words indicate the error of **TS**, orange words suggest the imperfection in **IW**, and green words are the problem of **TM** .

Example 1 and 2 are taken from the 92 documents in the testing data. They are translations of the same ST and their back-translations by Google MT engine. You may challenge the comparability between the ST and the TT in the first example. It is worth mentioning that when extracting the features for training, the researcher has normalised their occurrences either to ST length or TT length, and in comparison the log likelihood ratio statisics has been adopted to take into account the imbalance of the text sizes. In the meantime, we have to admit that translations by different

Example 1: Bad Translation:
Resistance and fear are generally the product of ignorance. Of course our hostility is no exception.
It is undeniable that when the bites of insects start to be unbearable, they become very annoying. When they spread disease, it is a threat, but emotionally, even harmful insects It still looks beautiful.
We have learned from the study of their remaining fossils that they have lived on the earth for nearly 400 million years.

抵制 和 恐惧 普遍 都 是 无知 的 产物 当然 我们 的 敌视 也 不 例外 。
无可否认 当 昆虫 咬伤 的 地方 开始 疼痛 难忍 的 时候 ， 它们 就 变得 非常 讨厌 ， 当 他们 传播 疾病 的 时候 ， 那 就是 一种 威胁 了 ， 但是 不带 感情 地 来讲 ， 即使 是 有害 的 昆虫 也 还是 长得 很漂亮 的 。
从 它们 的 遗留 化石 的 研究 中 我们 获悉 他们 已 在 地球 上 生存 近 4 亿 年 了 。

translators always vary in length, and it is a ubiquitous phenomenon that we evaluate translations of varying lengths to the same ST. Therefore, there is no reason to believe that all STs and TTs should be roughly of the same length for comparison. Looking at the feature items for both translations, we could identify a significant difference of feature distribution. Table 6.10 lists the log-likelihood ratio[4] of selected features in both examples (normalised). More stars indicate a higher significance (sig.) level. The statistical test reveals a tendency of differentiation between the 'good' (Example 2) and 'bad' translation (Example 1) in the occurrence of nouns, noun phrases, prepositional phrases, noun compounds and semantic roles (A0 and A2) so on. These log-likelihood statistics suggest that poor translations may have fewer such features, and in contrast they are more prevalent in good translations. This statistical test further confirms the previous feature selection in the last section.

In the following, I will discuss the feature presence in groups in the two examples for different quality aspects, as I cannot list all individual features. Example 1 is an incomplete translation and scored rather low by human annotators. As a large part of the ST has not been translated, the final scores for all the four quality aspects practically fall within their first quartiles. In the translation, we have identified errors in **TS**, i.e. 敌视 ('antagonism') in the first sentence and 生存 ('existence') in the third sentence. Each sentence contains one type of errors regarding the four quality aspects. In contrast, Example 2 is translated with fairly good quality. The translation contains 3-5 errors of each type in total. Therefore, it is scored in the third quartile of the specified ranges (in Table 5.11), and the translation is evaluated between 'acceptable' and 'strong' for most of those highlighted parts in Example 2 require only minimal or some post-editing work to be usable. For instance, sentence 2 in the translation contains a **UT** error and a **TM** error. We see that the erroneous sentence is caused by the translator's failure to interpret the coordinate clauses connected

[4] http://ucrel.lancs.ac.uk/llwizard.html

Example 2: Good Translation:
Everyone knows that hostility and fear are the products of ignorance, and our enemy is also no exception.
We have to admit that insects are really annoying when we bite our teeth and act as a threat to spread the disease. However, we calmly analyze that even toxic insects are beautiful.
From the perspective of fossils of insects studied, we know that insects have survived the Earth for 400 million years.
Today, insects can be seen everywhere.
Insects can be found either in the polar regions or on the tops of alpine mountains.
The bees, wasps, ants and termites each have some complex stories in the process of searching, defensive and breeding.
The working bees may even be assigned to different breeders throughout their lives. In the first three weeks, the young male bred the female bees and their bees, cleaned the bees, and swept their wings to make the entrance colder. Resist attack and siege by the invaders.
Only after this training is over, the bees are allowed to leave the cell, make nectar and spread pollen.
In addition to these deeds, some ants use leaf fragments as spoons to transport food to their ant nests. People often describe this behaviour of ants as "wisdom" and begin to compare insects with human society.
What saddens me is that people are reluctant to accept the fact that they can appreciate the simplicity and happiness of ordinary insects.
Discovering the structure and beauty of a new world only requires a magnifying glass for a short period of time.

SENT 1: 大家 都 知道 ， 敌视 和 恐惧 是 无知 的 产物 ， 而 我们 的 敌抗 也 是 不 例外 的 。

SENT 2: 我们 不得不 承认 昆虫 在 咬得 我们 发疼 及 作为 一种 威胁 传播 疾病 时 ， 的确 让 我们 感到 厌烦 。 但是 我们 冷静 地 分析 下 ， 即使 是 有毒 的 昆虫 ， 它 也 是 美丽 的 。

SENT 3: 从 研究 的 昆虫 遗体 化石 来看 ， 我们 知道 昆虫 已 在 地球 上 存活 了 4 亿年 了 。

SENT 4: 今天 ， 昆虫 到处 可见 。

SENT 5: 无论是 极地 之巅 还是 高山 之顶 都 能 发现 昆虫 。

SENT 6: 蜜蜂 ， 黄蜂 ， 蚂蚁 和 白蚁 ， 各自 在 寻食 ， 防御 和 繁衍 过程 中 ， 都 有 一些 复杂 的 故事 。

SENT 7: 工作 的 蜜蜂 一生 甚至 会 接连 分 到 不同 的 养殖 者 手中 ： 在 最初 的 三周 里 ， 年轻 的 雄蜂 饲养 着 雌蜂 和 它 的 蜂 蛋 ， 清洁 蜂窝 ， 用蜂翼 扇 着 ， 使得 入口 变冷 以 抵御 侵略者 的 袭击 和 围攻 。

SENT 8: 只有 在 这种 训练 结束 后 ， 小蜜蜂 才 允许 离开 蜂窝 ， 酿造 花蜜 和 传播 花粉 。

SENT 9: 除了 这些 事迹 ， 一些 蚂蚁 用 树叶 碎片 作为 匙 来 把 食物 运送 到 自己 的 蚁窝 ， 人们 往往 把 蚂蚁 的 这种 行为 描述 成 " 智慧 " ， 并且 开始 在 昆虫 和 人类 社会 中作 比较 。

SENT 10: 令 我 伤心 的 是 ， 人们 不愿 接受 能 从 欣赏 ， 普通 昆虫 的 精致 优美 上 获得 简单 快乐 这一 事实 。

SENT 11: 发现 一个 新世界 的 结构 和 美丽 只 需要 一小 段时间 的 放大镜 。

by ‘and’. The right understanding for this part is that we feel insects annoying since they bite us and we treat them as a threat as they spread disease. However, the translation for this part has mistakenly treated 昆虫 在 咬得 我们 发疼 (‘insects give us painful bites’) and 作为 一种 威胁 传播 疾病 时 (‘spreading disease as a threat’) as the reason for their annoyance. The translation for the highlighted parts in this sentence then can be improved if we render the coordinate clauses correctly. In addition, the segment highlighted in green in the middle 作为 一种 威胁 传播 疾病 时 is broken Chinese, which can be fixed by a slight reordering. I list the predictions by the models trained in this section and the original scores by human annotators in Table 6.11 for the two examples.

Translation	Quality	Human	QuEst	Full	TopNCorr	Kbest
Example 1	UT	5	6.07	5.41	4.11	4.69
	TS	3	5.10	6.07	3.81	4.03
	IW	4	5.76	7.15	5.11	4.72
	TM	3	3.42	4.23	3.77	3.98
	AD	8	11.76	10.99	7.71	10.06
	FL	7	9.36	9.29	8.38	8.33
	TO	15	22.96	20.91	20.60	15.14
Example 2	UT	21.5	24.17	21.47	21.97	21.82
	TS	17.5	16.45	15.74	16.48	16.60
	IW	19.5	16.95	16.49	16.76	17.01
	TM	9.5	9.95	9.78	9.22	9.64
	AD	39	40.67	36.55	37.44	38.72
	FL	29	24.51	25.59	27.06	26.72
	TO	68	71.11	63.57	61.63	66.53

Table 6.11 Model Predictions of Two Example Translations

As shown in Table 6.11 , models based on the proposed features, e.g. **Full**, **TopNCorr** and **Kbest** have improved estimations with smaller prediction errors for the 7 quality scores in comparison to the **QuEst** model. However, all models, either **QuEst** or **non-QuEst**, have yielded elevated estimations for low scores, showing a tendency of predicting higher than what they are (e.g. Example 1). For higher scores, e.g. Example 2, the three models trained with the proposed features predict them more accurately with closer estimation.

The partially superior performances of both **Kbest** and **TopNCorr** together have demonstrated that more predictive models for most quality aspects can be built with this feature set over the baseline features to estimate different quality component scores. Different feature selections have shown models built with the selected features can achieve satisfactory performance with the current grid search setting.

Models	Target	MSE	r	ρ	τ
QuEst	UT	38.82	0.57	0.54	0.37
	TS	34.21	0.42	0.48	0.36
	IW	13.75	0.5	0.56	0.4
	TM	13.6	0.29	0.35	0.26
	AD	58.34	0.63	0.65	0.48
	FL	36.68	0.35	0.37	0.27
	TO	121.93	0.65	0.69	0.51
Full	UT	40.18	0.6	0.57	0.43
	TS	37.35	0.58	0.56	0.42
	IW	19.31	0.43	0.38	0.26
	TM	4.67	0.64	0.6	0.47
	AD	90.97	0.68	0.69	0.56
	FL	29.46	0.61	0.57	0.41
	TO	122.95	0.63	0.62	0.53
TopNCorr	UT	34.59	0.55	0.58	0.43
	TS	30.85	0.52	0.52	0.39
	IW	12.39	0.57	0.54	0.39
	TM	10.84	0.49	0.49	0.39
	AD	46.57	0.59	0.62	0.46
	FL	49.79	0.62	0.64	0.45
	TO	117.41	0.6	0.66	0.53
Kbest	UT	30.17	0.61	0.62	0.45
	TS	25.73	0.63	0.62	0.41
	IW	21.58	0.58	0.46	0.32
	TM	15.18	0.49	0.43	0.28
	AD	43.09	0.71	0.69	0.55
	FL	26.54	0.59	0.65	0.48
	TO	103.46	0.64	0.67	0.50

Table 6.12 Feature-based Models on MT Data

6.4.4 Application of Document Level MTQE

On the MT data annotated with translation errors in Chapter 5, I compare the **QuEst** and other three models such as **Kbest**, **TopNCorr**, and **Full** to see how these models with different subsets of feature perform on this MT data. To this end, I extract the same number of features for each model as Table 6.9, and apply the four trained models above directly to evaluating these MT data. The results are given in Table 6.12. As shown in the table, **QuEst** predicts **UT**, **TM** and **FL** less satisfactorily with rather lower correlations (e.g. Pearson Correlation r) with human annotation on these aspects. However, it has achieved the highest correlation in estimating **TO** scores. In contrast, **Full** seems to improve on all aspects except that it is outperformed by **QuEst** with respect to prediction for the **TO** scores. **TopNCorr** demonstrates a rather stable but mediocre performance in comparison to the former two models. However, it has achieved the best correlation in predicting **FL** among

the four models. **Kbest** has shown better correlations with human judgements in most quality aspects.

6.5 Summary

I treat the document-level quality estimation as a supervised machine learning task, in which a series of models are trained on a collection of manually annotated documents and applied to estimate the quality of new translations.

Overall, models trained with feature selected using variable ranking (i.e. by the criterion of importance to the model using 'k-best' method) have demonstrated the potential to estimate with reasonably better correlation with human judgements and reduced MSEs, in comparison to models trained with the full feature set and selected features by the correlation threshold. With current feature set and annotation scheme, we could build fine-grained quality estimation models to predict new, unseen translation. The models discussed in previous sections have been trained from HT data and then applied to MT data. Different models show their potential strengths in estimating various aspects of quality scores. In general, models trained with the proposed feature set correlate better with human judgements and outperform the model built with the baseline feature set. Capable of fine-grained quality estimation, this method is advantageous for offering fast quality feedback from multiple aspects, providing more insight into the fine-grained translation quality.

Chapter 7

Deep Learning-based Sentence-level HTQE

7.1 Introduction

In the last chapter, I have dealt with HTQE at the document-level, and in this chapter, I propose a deep learning-based model for HTQE at the sentence level.

As I previously discussed in the chapter of introduction, HTs are often evaluated at the document level for a summative assessment and the segment (e.g. word, phrase, sentence) level for a diagnostic purpose. At a macro level, quality labels, continuous or categorical, are sought as an overall evaluation. In contrast, sometimes translations need to be evaluated at a micro level for more instructive feedback and other applications (e.g. error analysis, post-editing), based on the results of the diagnostic analysis. Thus, the segment-level, particularly sentence-level TQE is complementary to document-level TQE.

Discrete manual features have been studied for many NLP tasks. Most reference-free methods are based on effective feature templates as I have discussed in Chapter 2 and 3. However, these features tend to be complex and require extensive feature engineering (Gupta et al., 2015). Linguistic processing tools and resources, such as constituency parser, discourse parser and large parallel data, are compulsory but not necessarily available for the language pair under examination. Most importantly, feature-based methods for sentence-level QE often suffer from the data sparsity issue at the segment level, as I discussed in Chapter 5. Therefore, it is considered essential to look for an alternative solution to HTQE at the sentence level.

Compared with discrete models with manual indicator features, neural network models are advantageous in two-fold. First, neural network models take low-dimensional dense embeddings (Collobert et al., 2011; Mikolov et al., 2013a; Pennington et al., 2014) as the input, which can be trained from a large-scale dataset, thereby overcoming the issue of sparsity. Second, the resulting neural features can capture complex non-local syntactic and semantic information that discrete indicator features can hardly encode.

Another motivation for applying deep neural networks to the sentence-level HTQE is that I have access to several thousand pairs of source-target sentences, which are often mandatory for training effective neural models. In this sense, the shortage of the document-level data hinders exploring the performance of the neural networks models to predict the quality of document-level translations. Neural networks have also been noted incapable of modelling long documents that have a hierarchical structure and comprise sentences of differential informativeness, i.e. complex semantic and syntactic structures (Tang et al., 2015; Liu et al., 2018).

For the above-mentioned reasons, I argue that deep learning-based methods making use of vector space models (Mikolov et al., 2013c) embed word vectors in a continuous vector space containing semantic and contextual information, without much fuss of feature engineering and reliance on language-dependent resources. Such information helps to overcome the shortcomings of losing semantics and feature sparsity characteristic of conventional manual feature engineering methods for short texts. Thus, it is worthy of exploring their application to the sentence-level HTQE.

This chapter presents a novel hierarchical neural network model (NeuralTQE) for estimating the fine-grained human translation quality at the sentence level. The contribution of the current study is fourfold. First, this work investigates the effectiveness of neural networks based learning to predict the fine-grained scores on the manually annotated HT data. Second, extensible to different language pairs, the proposed framework requires only pretrained word embeddings and no extra feature selection that is common to the conventional feature-based methods. Also, I applied the method to MTQE data, researching and showing its applicability to task-specific QE for HTs and MTs. In addition, the current approach has introduced a novel compact neural network with attention mechanism in order to capture both local and global information. This HTQE implementation will be an extension to the existing neural networks based models for MTQE and those conventional ones as well.

7.2 Related Work

Recent years have seen a boosted interest in using the deep learning approach for virtually all areas of NLP, particularly in MTQE that is related to my work.

Gupta et al. (2015) presented a compact MT evaluation based on the Tree Long Short Term Memory (Tree-LSTM) networks (Tai et al., 2015). The proposed metric makes use of *glove* word vectors (Pennington et al., 2014) and the dependency Tree-LSTM neural network implementation to rank the MT translations by computing their similarity with human references on an automatically converted training data. Their method has obtained a system-level (normalised aggregation of segment scores) correlation above 0.9 for MT translations of four language-pairs and a maximum

0.438 segment-level correlation. Though the method is competitive to the current complex alternative approaches to MT evaluation that 'involve system combination, extensive external resources, feature engineering and tuning' (Gupta et al., 2015), it presupposes a parallel set of gold-standard references like any other conventional reference-based approaches.

Guzmán et al. (2015) introduced a simple feed-forward neural networks based framework for pairwise MT evaluation that aims to select the better translation from a pair of hypotheses, given the reference translation. They use distributed vector representations of the translation and references and feed them into a multi-layer neural network that models the interaction between each reference and candidate translations in a pairwise setting. Their method integrates several layers (e.g. semantic vectors of words, syntactic vectors of sentences) of linguistic information, including external features (e.g. the BLEU scores of the translations) with those about both the references and the two alternative translations simultaneously in a simple feed-forward neural network learning architecture. In their study, when vectors from word embeddings are used in combination with four commonly-used metrics (BLEU, NIST, TER and METEOR) (Doddington, 2002; Papineni et al., 2002; Banerjee and Lavie, 2005; Snover et al., 2006), the system achieved significant improvements, offering generalizations that complement very strong metric combinations, and claimed to be able to train an optimised task-specific cost function for the pairwise MT evaluation setting. However, the framework also depends on references and does not take into consideration the source side information which is intuitively crucial for the translation process.

Paetzold and Specia (2016) proposed the SimpleNets approach for sentence-level QE. The solution was originally advanced to assess text simplification quality, for which SimpleNets consists of five steps:

- decomposition of the original and simplified sentences into n-grams of a maximum size M
- obtaining the union of n-grams from the original and simplified sentences
- assigning quality labels to the n-grams in the pools according to the principle of compositionality
- representing the sentences with n-grams in the form of continuous vectors
- Long Short-Term Network Memory (LSMT) training instances in minibatches

The quality of all n-grams is then merged using a certain policy such as averaging to produce the quality at the sentence level. However, for QE purpose, they look at only one side n-grams of the translation pairs on the hypothesis that sentence-level QE can be learnt from either source-side or target-side information. Thus, they train two variants of SimpleNets, i.e. source sentence-based or target sentence-based, with only some small adaptations to the original framework such as replacing its

softmax activation modes with a single dense node and the cross-entropy loss function with Mean Average Error (MAE) function for regression.

Kim and Lee (2016a) put forward a recurrent neural networks (RNN) method for QE at the sentence level. Their proposal is basically to extract the last layer of RNN containing the information about how well a target word is translated from an ST. They use gated hidden unit (Cho et al., 2014) as an activation function to learn the long-term dependencies of translation quality for target words. A logistic sigmoid function is then used to compute the quality score (HTER) from the last hidden state as the summary unit of condensed quality vectors. They extended this model in Kim and Lee (2016b) to the word- and phrase-level QE, using the concatenated and averaged hidden states of backward and forward vectors.

In contrast to the afore-mentioned neural networks based methods, the model I proposed differs in the sense that it dos not focus on comparing the similarity between the gold-standard reference and the target translations with either a purpose of pairwise selection or reranking translations at the system and segment level. These models discussed rely solely on variants of the LSTM architecture. In contrast, the proposed learning architecture combines the strength of both convolution neural networks (CNN) and RNN with an additional cross attention mechanism that adds the ability to capture long-range dependencies. Most importantly, the goal of this research is to predict fine-grained direct assessment quality scores (Graham et al., 2017a) for human translations by professional evaluators, an apparently more challenging task (Guzmán et al., 2017) than learning the relative ranking of translations or estimating the similarity between candidate translations and references. Additionally, I attempt to extend this framework to task-oriented QE on MTQE data, using only dense vector representations.

7.3 Models

In this section, I introduce the components of my neural network architecture as shown in Figure 7.1. Given a translation pair, the source sentence x and the target sentence y are encoded into a fixed-sized vector representation through two separate CNN-BiLSTM-Attention architectures. Denoting the final vectors as $\mathbf{x}$ and $\mathbf{y}$ respectively, the proposed model predicts seven quality scores (**UT**, **TS**, **IW**, **TM**, **AD**, **FL** and **TO**) using a linear regression based on the concatenation of $\mathbf{x}$ and $\mathbf{y}$. I describe the neural components in a bottom-up order.

7.3.1 Context-aware Word Representation

Given a source sentence x or a translation y, which can be represented by $w_1, w_2, \ldots, w_n$, we transform the words into vector representations. While this could be simply achieved by a word embeddings layer (Bengio et al., 2003), words are

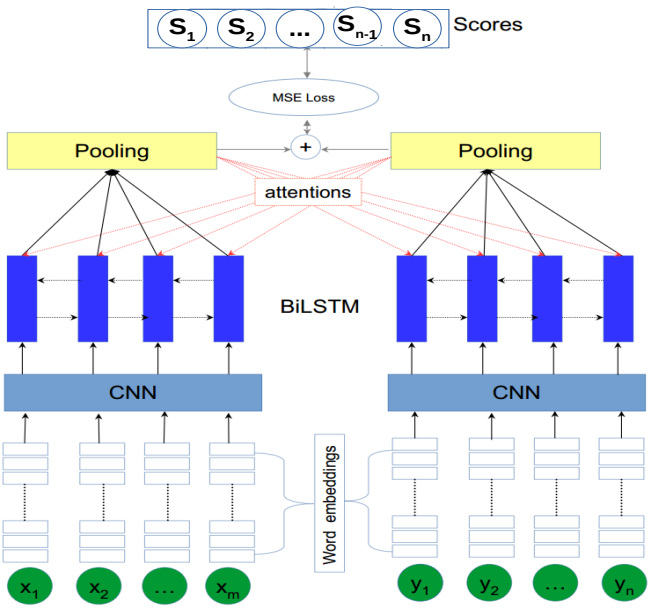


Figure 7.1 Neural Model Structure

often ambiguous unless they are given within a specific context. To obtain more precise word-level information, I build multiple convolution layers upon the standard word embedding layers for context-aware word representation. The reason I choose the CNNs as a feature extractor has been highlighted by Goldberg and Hirst (2017, pp.152) that:

> the CNN is in essence a feature-extracting architecture. It does not constitute a standalone, useful network on its own, but rather is meant to be integrated into a larger network, and to be trained to work in tandem with it in order to produce an end result. The CNN layer's responsibility is to extract meaningful sub-structures that are useful for the overall prediction task at hand.

For the convolution layer of a width k, I apply multiple kernels $\mathbf{H}_\mathrm{i} \in \mathbb{R}^{\mathrm{d}\times(2\mathrm{k}+1)}$ before a non-linearity transformation. Specifically, for a window centred at i-th word, the output $\mathbf{f}_\mathrm{i}$ is given by:

$$\mathbf{f}_\mathrm{i} = \mathrm{relu}(\langle \mathbf{H}_\mathrm{i}, \boldsymbol{w}_{[\mathrm{i}-\mathrm{k}:\mathrm{i}+\mathrm{k}]} \rangle + \mathrm{b}_\mathrm{i}),$$

where $\boldsymbol{w}_{[\mathrm{i}-\mathrm{k}:\mathrm{i}+\mathrm{k}]}$ denotes the window size, and b_i a bias. The word representation is then the concatenation of all convolution layers.

7.3.2 Sentence-level Representation

To capture the global information of a sentence, bidirectional LSTMs (Graves et al., 2013) are used on $\mathbf{f}_\mathrm{i}$. The outputs include a sequence of forward hidden states $\overrightarrow{\mathrm{h}}_1, \overrightarrow{\mathrm{h}}_2, \ldots, \overrightarrow{\mathrm{h}}_\mathrm{n}$ and a sequence of backward hidden states $\overleftarrow{\mathrm{h}}_1, \overleftarrow{\mathrm{h}}_2, \ldots, \overleftarrow{\mathrm{h}}_\mathrm{n}$. One takes the input word sequence in its original order and the other takes the sequence

in the reverse order. The two sequences are then concatenated into one $h_i = \overrightarrow{h_i} || \overleftarrow{h_i}$. In this way, each annotation h_i contains the summarized information about the whole input sentence, but with a strong attention to the details surrounding the i-th word.

7.3.3 Attentive Reading

In NLP, translation is a task of modelling a pair of sentences. The task itself correlates with other NLP tasks, such as answer selection (Yu et al., 2014), paraphrase identification and textual entailment (Bowman et al., 2015). Yin et al. (2016) argue that most prior work in these tasks, relying on manually designed, task-specific linguistic features, models each sentence's representation separately and rarely considering the impact of the other sentence, they propose to model a pair of sentences with the attention-based convolutional neural network. In line with their work, I consider it important to consider the mutual influence of the source target sentence pair in the context of translation. It conforms to what humans do when translating a sentence from one language to another. Two sentences are rarely processed independently of each other. It is necessary to value this interdependency when we are modelling the behaviour of human translation.

In the meantime, different parts of a pair of sentences contribute unequally to the semantic adequacy and language fluency of the final output. When translating, translators usually focus on specific parts of the two sentences and read back and forth to come up with an optimal translation within their capacity. Segments of the two sentences are related to each other with different degrees of correspondence.

In this process of composing the new sentence, a critical aspect of translation every translator encounters is that we need to refer back to the original text, focusing on specific parts that are relevant to the proposed translation. At each step, we should pay attention to the most relevant parts of the source text and the partially completed translation so that we can make an optimal decision about the next word to choose. This phenomenon of attention has proved useful in machine translation (Bahdanau et al., 2015; Luong et al., 2015b).

After obtaining the sentence representations centred on different words, I simulate the repeated reading and aligning process of human translators and design a cross attention mechanism to pinpoint the information particularly important for quality estimation. This process is called as distillation via attentive reading. The approach closely resembles the method introduced in Buduma and Lacascio (2017, pp.197), except that I compute the attention score differently. First, I create a scalar (i.e. a single number) as the relevance score for each of the outputs from the convolutional layer. The score is computed in the form of a linear transformation between each encoded output and the decoded state at time stamp t-1 (i.e. a previous step). These scores are then normalised using a softmax operation (see Equation 7.1) to be in the range $(0, 1)$, and they are further used to individually

scale the representations of either source sentences and target sentences from the convolutional layer before these representations are plugged into the concatenation operation. The idea behind this method is that these weighted scores signify how important individual input representations are to the decision for the output decoding at time step t.

$$softmax(x) = \frac{exp(x_i)}{\sum_j exp(x_j)} \tag{7.1}$$

Specifically, in my implementation, at each time step i, the model infers a variable-length weight vector α_i based on the current hidden states of source side ($\mathbf{h}$) and target side ($\mathbf{h}_i$) output. A weighted average of the source representations is then computed as the global context vector over all the source states to decide which parts of the translated sentence are important for quality estimation and vice versa. Given h_i for each word, the final sentence representation after the attention mechanism then is:

$$\mathbf{s} = \sum_i^n \alpha_i h_i,$$

where α_i is the attention weight for $\mathbf{h}_i$ and it is computed by:

$$\alpha_i = \frac{\exp(f(\mathbf{h}_i, \mathbf{h}))}{\sum_i^n \exp(f(\mathbf{h}_i, \mathbf{h}))},$$

with $\mathbf{h}$ being the average of its counterpart. The above two equations are a weighted sum layer and a softmax layer computing probabilities. Their task is to make a weighted sum to address the issue of knowing which positions are more relevant when trying to predict next position's output. The score function f is:

$$f(\mathbf{h}, \mathbf{h_i}) = \boldsymbol{v}^T \tanh(\boldsymbol{W}_{a1}\overline{\mathbf{h}} + \boldsymbol{W}_{a2}\mathbf{h}_i),$$

where $\boldsymbol{v} \in \mathbb{R}^{d_a}$, $\boldsymbol{W}_{a1} \in \mathbb{R}^{d_a \times 2h}$ and $\boldsymbol{W}_{a2} \in \mathbb{R}^{d_a \times 2h}$ are trainable parameters.

7.3.4 Training

Given a training triple (x, y, s), where x is the source sentence, y is the translated sentence and $s \in \mathbb{R}^k$ is the score vector annotated by human judges from k different aspects, respectively. MSE loss is used for training.

$$\ell(x, y, s) = \frac{1}{k} \sum |\text{SCORE}_i(x, y) - s_i|^2 + \lambda ||\Theta||^2$$

I use a stochastic gradient method to train the parameters. For optimisation, I choose Adam, an algorithm 'for first-order gradient-based optimisation of stochastic objective function' that is memory efficient and 'well suited for problems that are large in terms of data and/or parameters' (Kingma and Ba, 2014, pp.1). The method computes adaptive learning rates for different parameters from the estimates of first

	# sentence	average length (words)	
		EN	ZH
Training	3000	24.48	24.99
Test	529	21.47	25.65

Table 7.1 Statistics of Translational Sentences

and second moments of the gradients, combining two recently popular methods, namely AdaGrad (Duchi et al., 2011) and RMSProp (Tieleman and Hinton, 2012). It has been known for the advantages of invariant magnitudes of parameter updates, non-stationary objective and adaptive to sparse gradients step size annealing (Kingma and Ba, 2014).

To avoid over-fitting, I apply a dropout operation (Srivastava et al., 2014) to each layer. The key concept of randomly dropping out units from the neural network during training is beneficial in preventing units from over co-adaptation (Hinton et al., 2012). In other words, dropout randomly zeros some of the elements of the input with a threshold probability p. This effective technique trims the neural networks in order to reduce overfitting.

Perhaps the most similar method to ours is Kim and Lee (2016a) except that the LSTM layer in this research is stacked on the hidden states generated through the cross-lingual attention mechanism on top of a convoluted layer using word representations as inputs, and they learn representations from large parallel corpora instead.

7.4 Experiments

I conduct a set of experiments on the sentence level with a corpus of trainee translation data.

7.4.1 Data

The translational data used for the training and testing in this study come from the published Parallel Corpus of Chinese EFL Learners (Wen and Wang, 2008). As introduced in Chapter 5, sentences from 50 translational documents for each source text are selected and annotated by two annotators according to the same annotation guideline ATA rubric. Therefore, I have 3529 English-Chinese translation pairs,

All the sentences have been scored per the same quality annotation scheme for the translation documents, as described in Chapter 5.

7.4.2 Setup

For evaluation, similar as in Chapter 6, I use the Pearson Correlation Coefficient (r), Spearman's rank correlation coefficient (ρ) and Kendall's tau (τ) to measure the association between model predictions and human judgements.

I use the pre-trained word embeddings to initialise the word representations. For English, the pre-trained 200 dimension GloVe vectors (Pennington et al., 2014) are used. For Chinese, I train 200-dimension word embeddings with Chinese Wikipedia dump[1], using Gensim package (Řehůřek and Sojka, 2010) with default settings. For other languages, including Czech, German (in Section 7.7), I use the pretrained word embeddings by fastText (Bojanowski et al., 2016).

For comparison, I also implemented the four feature-based models **QuEst**, **Full**, **TopNCorr** and **Kbest**, using the XGBoost algorithm with grid search. I use the full set of sentence-level features for **Full**, and select the same subsets of features for **TopNCorr** and **Kbest** as in Chapter 6 for the reason of convenience. Note that I excluded a set of cohesion and coherence features that do not exist for translation sentences[2]. The performance of these feature-based models are compared to the neural model with respect to fine-grained sentence-level HTQE.

7.5 Results and Discussion

Table 7.3 presents the results of the proposed models. To show the effectiveness of the attention mechanism, I implemented a model without the attention (w/o). Table 7.2 lists the hyper-parameter settings of the final models. The lowest correlations with human judgements for each quality score in all the models are highlighted in reddish colour, and the highest are highlighted in green. The neural models rank the second best regarding correlation and are highlighted with lighter green.

word embedding size	$\mathrm{d} = 200$
window size	$\mathrm{k} = [1, 2, 3, 4]$
Initial learning rate	$\alpha = 0.001$
dropout rate	$\mathrm{p} = 0.5$
regularization	$\lambda = 1e-3$
number of layer	1

Table 7.2 Hyper-parameter settings

From the perspective of correlation with human judgements, feature-based **TopNCorr** models, which are built from the top correlative features with each quality aspects as I have discussed in Chapter 6, have outperformed all other models significantly in predicting **UT**, **TS**, **AD** and **TO**, with nearly 20+$\%$ higher correlation

[1] https://dumps.wikimedia.org/zhwiki/20171103/zhwiki-20171103-pages-articles-multistream.xml.bz2

[2] For the detailed list, refer to Appendix D

Model	Target	MSE	r	ρ	τ
Feature-based Model					
QuEst	UT	31.01	0.37	0.37	0.26
	TS	11.20	0.35	0.34	0.24
	IW	4.57	0.24	0.25	0.17
	TM	3.70	0.25	0.25	0.18
	AD	74.75	0.39	0.38	0.27
	FL	18.97	0.24	0.25	0.16
	TO	155.92	0.36	0.36	0.25
Full	UT	31.32	0.33	0.34	0.23
	TS	13.22	0.24	0.26	0.19
	IW	4.86	0.20	0.20	0.15
	TM	3.64	0.25	0.23	0.16
	AD	80.21	0.32	0.33	0.22
	FL	18.21	0.27	0.27	0.19
	TO	155.80	0.33	0.33	0.24
TopNCorr	UT	37.72	0.54	0.53	0.38
	TS	15.25	0.55	0.56	0.40
	IW	8.73	0.33	0.38	0.26
	TM	4.71	0.35	0.37	0.26
	AD	96.92	0.55	0.55	0.39
	FL	24.66	0.34	0.36	0.25
	TO	180.99	0.53	0.53	0.38
Kbest	UT	53.17	0.39	0.38	0.31
	TS	21.61	0.27	0.30	0.21
	IW	9.74	0.18	0.19	0.13
	TM	5.30	0.18	0.20	0.14
	AD	139.87	0.29	0.31	0.21
	FL	28.52	0.18	0.20	0.14
	TO	264.24	0.29	0.30	0.21
NeuralTQE					
w/ attention	UT	36.49	0.41	0.42	0.29
	TS	14.13	0.38	0.41	0.29
	IW	7.16	0.31	0.30	0.22
	TM	4.16	0.35	0.35	0.22
	AD	94.55	0.40	0.42	0.29
	FL	18.02	0.32	0.34	0.24
	TO	176.64	0.42	0.43	0.30
w/o attention	UT	64.41	0.22	0.23	0.16
	TS	25.65	0.22	0.23	0.15
	IW	11.46	0.13	0.13	0.09
	TM	5.45	0.16	0.16	0.11
	AD	168.90	0.23	0.24	0.16
	FL	28.98	0.15	0.15	0.09
	TO	296.63	0.22	0.23	0.09

Table 7.3 Sentence-level HTQE results

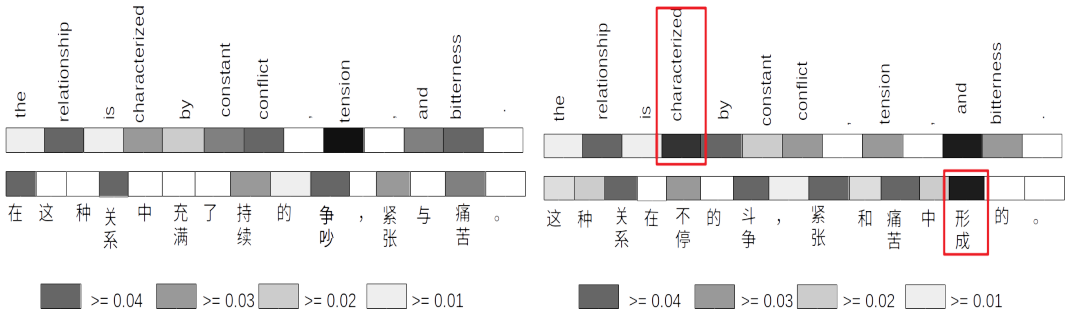
with human evaluation than **QuEst**, **Full**, **Kbest** and **NeuralTQE** (w/o attention). Welch's t test shows that: **QuEst** is almost as good as **Full** (t = 1.8969, df = 33.077, p > 0.05), which is consequently no better than **Kbest** (t = 0.60784, df = 37.867, p > 0.05). While the **NeuralTQE** with attention mechanism demonstrates better estimation than **QuEst** (t = 2.0446, df = 35.326, p < 0.05) and **Full** (t = 4.3977, df = 39.325, p < 0.05). **NeuralTQE** without attention mechanism then is worse than every other model, including **Kbest** (t = 3.4492, df = 35.352, p < 0.05). Overall, **NeuralTQE** with attention mechanism ranks the second best, outperformed by **TopNCorr** in several aspects (**UT**, **TS**, and **AD**) by a margin of approximately 15% higher correlation (t = 2.8719, df = 33.814, p < 0.05). However, **NeuralTQE** with attention mechanism does perform on par with **ToNCorr** in predicting **IW**,**TM** and **FL** (t = 1.1189, df= 15.965, p > 0.05).

As I have previously discussed in Chapter 5, **IW** and **TM** are more related to language use and conventions and have been collapsed to **FL**. The findings above suggest that predicting **FL** is more challenging than predicting **AD** that includes **UT** and **TS**, as shown by the correlation scores in the shaded cells in red. The stacked neural model with attention mechanism has manifested the potential of estimating translation quality with respect to language use in particular. It shows the effectiveness of a deep learning-based model without manual engineering and reliance on other language-dependent resources such as parsers, POS taggers and SRL labellers. Using word embeddings as representations of STs and TTs, the proposed neural architecture prove its capability of capturing the short- and long-range dependency that matters to quality between them.

In comparison to the MTQE based **QuEst** baseline features, **TopNCorr**, the feature-based model built on subsets of the proposed features, has achieved an overwhelming performance, showing its applicability to both document-level and sentence-level HTQE.

In contrast, the neural model with attention has achieved the second place in term of correlation with human judgements (Pearson as the primary comparison). Nevertheless, without the attention mechanism, it is no better than the worst **Kbest** model in the experiment, implying that the attention mechanism helps boost the neural model's strength of leveraging the automatically learned semantic and syntactic information from the pretrained word embeddings.

Kbest, best for the document-level HTQE, suffered a significant performance deterioration for the sentence-level HTQE. At the sentence-level, features selected on the F statistics criterion, such as source target conjunct log ratio and phrase alignment information, become very sparse and introduce noise to the prediction task. The inclusion of such sparse features could be problematic for estimation.



(a) Good Translation (b) Bad Translation

Figure 7.2 Attention Plot

7.6 Case Study

7.6.1 Attention Visualisation

In view of the importance of attention mechanism in my implementation to model the source-target sentence pair, I visualise the attention weights extracted from the training process for a pair of translation sentences, as shown in Figure 7.2.

In the figure, the progressive bar of greyness (from light white to dark black) indicates the weight of individual elements in the output representation at time step t. Figure 7.2a plots the attention weights for the words in a good translation (UT 23.5, TS 20, IW 19, TM 11), and Figure 7.2b in a bad translation (UT 10, TS 9.5, IW 18, TM 10.5). It is found that

- **More attention weights are given to the content words with variation.** The attention mechanism in the sequence to sequence neural machine translation models focuses on 'local' instead of purely 'global' attention, and as a consequence it is more like 'local alignment' process to achieve either monotonic or predictive alignment (Luong et al., 2015b). In contrast, the attention mechanism in my approach, as manifested by the dynamic weights, does not seek an alignment. For each sentence pair (i.e. ST and TT), higher weights are given to content words discriminatingly in both ST and TT for either good or bad translation. For instance, the Chinese word 紧张 ('tension') is less weighted than its correspondence 'tension' in English, and both the English verb 'characterize' and its translation 充满 ('full of') are not selected as important elements for the good translation in Figure 7.2a.

- **Attention weights also highlight potential translation errors.** In the bad translation, 'characterized by' and the 'and' in the coordinating conjunction have been assigned higher weights than other words in the ST. The corresponding words in the TT again do not acquire the same attention weights. Nevertheless, the wrong translation 形成 ('form') of 'characterized by' has been notably assigned the highest weight. This phenomenon suggests the 'global' nature of the attention mechanism has caused the dynamic weights of the words in STs

and TTs, and the attention itself tries to highlight segments in the sentence that differs from others. Among the differences between the TTs, some are correct translations, and some are not. Specifically, in the TTs, the highlighted parts are often important features (e.g. noun phrases, verb phrases) as discussed in Section 5.4.2 and 6.4.2, including those that contain potential errors (e.g. 形成).

7.6.2 Error Analysis

Turn now to the evidence on the accuracy of the neural models obtained to predict the four component quality scores[3]. Figure 7.3 provides an overview of the distribution of the human annotated scores and the predicted scores by the proposed models.

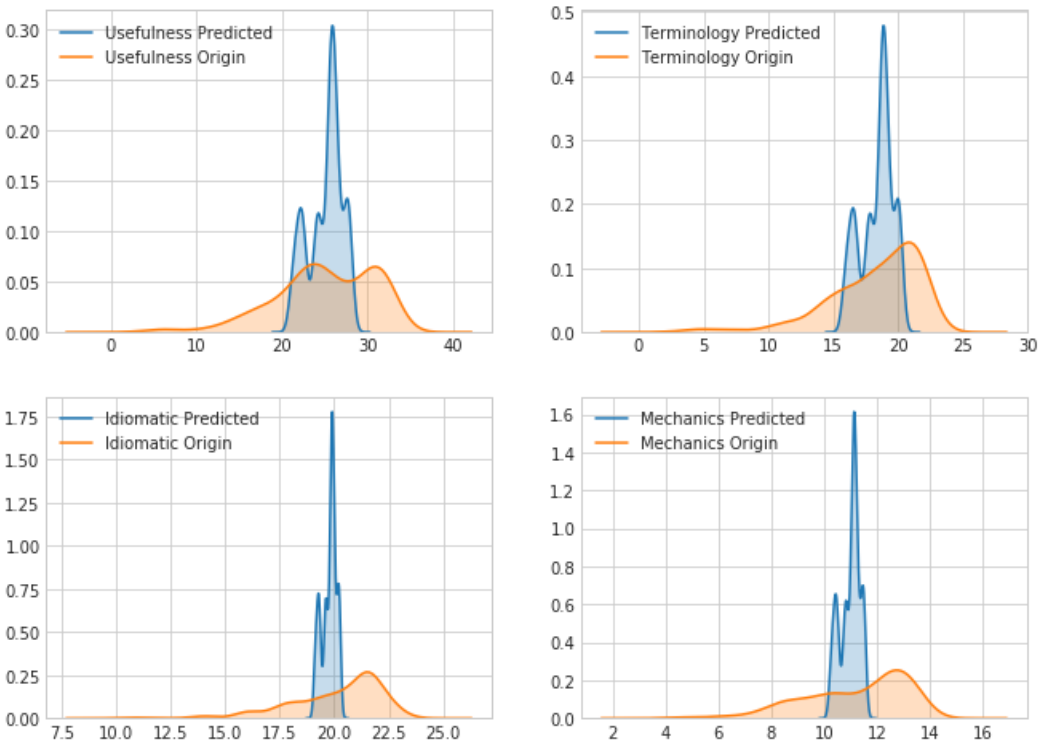


Figure 7.3 Distribution of Human Judgements and Model Predictions

As Figure 7.3 shows, there is a clear trend of contracting towards the third quartile of the original scores (human annotations) for each type of translation quality scores. This tendency suggests that the proposed models may have deteriorating performance in estimating extreme scores on the continuum. In other words, the trained models may predict higher the first quartile scores and lower the fourth quartile scores with more significant variance. This assumption warrants further statistical proof, though. In the following, I will discuss this phenomenon with some specific instances.

[3]As **AD**, **FL** and **TO** are mostly a summation of weighted summation of these component scores, and their distribution and estimation strongly correlate with these subscores as shown in Table 7.3.

The upper example in Table 7.4 contains English source sentences and its corresponding Chinese translations. The differences between the human scores and model estimations are huge. A closer examination of the translations reveals that the translation has twisted the meanings of the source sentences due to mistranslations of the English word 'surgeon' as 手术 ('surgery'). Also, the Chinese word 寄托 ('place hope on'), which does not exist in the original, has changed the meaning of the translation. As a consequence, the whole sentence needs to be re-translated, which explains the low scores by human annotators for **UT** and **TS**.

I suspect that features in the proposed framework cope with translation quality at a somewhat abstract level, and as a consequence, they are insensitive to the changes of meaning. As for the proposed neural model based on word representations, it may be biased towards word level adequacy such that significant changes of meaning due to addition, untranslation and mistranslation to close synonyms could not be detected accurately (in the case of the upper example). Such semantic intricacy requires a model to better capture the underlying meaning of sentences for those underscored 'good' translations. It is the same case with the lower example, in which 更多 的 目光 投向 ('set eyes on') is a non-literal but valid translation for 'renewed interest in'. Thus, freer but still valid translations do pose a challenge to the models as well (in the case of the lower example).

Another possible explanation is that the trained models are tuned to minimise the MSE and thus tend to predict worse the quality in the tails of the quality range. This finding corroborates with the findings in Moreau and Vogel (2014). They found in MT quality estimation the issue of the tails prediction problem generally holds true for all datasets and systems, particularly for supervised methods. Optimisation in supervised systems has intrinsic flaws in assessing the 'very good' and the 'very bad' translations.

In these two specific examples, all models, including the best feature-based **TopNCorr** and the second best neural model **NeuralTQE**, have shown the tendency of augmenting the estimations for low-quality translations and underscoring for very high-quality translations. Notwithstanding the neural model predicts best among all models for the very low and very high **UT** and **TS** in such two cases (closer to human annotations), the differences between its estimations and human annotation are huge. The relatively better scores of the proposed neural model as compared to the feature-based methods for **UT** and **TS** suggest that a neural network is more capable of modelling semantic adequacy, compared to conventional feature-based methods.

7.7 Task-oriented MTQE with NeuralTQE

HTER Score Prediction

It would be interesting to apply the sentence-level HTQE framework to MTQE data.

	Model	UT	TS	IW	TM	AD	FL	TO
Freedom from this constraint is the dream of every transplant surgeon . 打破 这 种 局限性 的 梦想 就 寄托 在 了 每次 移植 手术 上 了 。	Human	6	4.5	21.5	12.5	10.5	34	44.5
	QuEst	23.3	16.4	18.0	9.2	43.2	31.5	69.7
	Full	17.7	13.7	16.7	9.4	38.2	29.5	71.6
	TopNCorr	15.9	13.3	**21.4**	**11.6**	**11.9**	28.8	**62.1**
	Kbest	28.3	23.7	26.5	19.4	48.3	29.5	71.6
	NeuralTQE	**12.9**	**10.5**	**20.2**	**11.5**	15.5	**31.1**	71.5
So far attempts to make artificial organs have been disappointing :nature is hard to mimic. hence the renewed interest in trying to use organs from animals . 到 目前 为止 ， 尝试 模拟 人造 器官 的 结果 让 人 颇 有些 失望 ： 自然 难以 模拟 。 因而 人们 将 更 多 的 目光 投向 动物 的 器官 上 。	Human	33.5	22.5	22.5	13.5	56	35.5	91.5
	QuEst	23.9	15.9	17.3	11.3	43.6	31.9	76.1
	Full	28.9	19.5	21.1	11.9	49.3	30.6	82.7
	TopNCorr	25.1	16.7	20.9	13.4	56.1	36.22	82.5
	Kbest	19.4	15.1	18.4	11.2	34.5	29.6	63.9
	NeuralTQE	**29.6**	**20.9**	18.5	10.5	36.5	29.1	85.5

Table 7.4 Human Annotation and Model Predictions

	DE-EN		EN-DE	
	sentences	words	sentences	words
training	23000	404,198	25000	453666
development	1000	19487	1000	18152
test	2000	35577	2000	36119

Table 7.5 Staistics of WMT17 dataset

Therefore, I apply this stacked neural solution to the sentence-level MTQE task which scores translation sentences according to the proportion of words that need to be changed, using HTER (Snover et al., 2006) as a quality score (Specia et al., 2009a). I did not implement the feature-based methods in this task, considering the time constraint and the complexity of preparing the necessary linguistic resources, in addition to the lack of knowledge of these languages.

I use the released German-English (DE-EN) and English-German (EN-DE) datasets (Specia and Logacheva, 2017) in the WMT17 quality estimation shared task[4], comparing the neural models with other participating systems in the shared task. The statistics of the dataset can be found in Table 7.7.

For evaluation, I use Pearson's correlation r as the primary metric as it was in the official ranking of the WMT17 shared task (Bojar et al., 2017). Results are shown in Table 7.6. All the results are obtained with the same configuration as in Section 7.5 for HTQE.

Results show that the NeuralTQE model shows no significant improvement over the baseline, suggesting much room to improve for the model to be comparable to the winning systems.

[4]http://www.statmt.org/wmt17/quality-estimation-task.html

	r	
Model	**DE-EN**	**EN-DE**
WMT17 baseline	0.45	0.39
WMT17 1st Winner (Kim et al., 2017)	0.73	0.69
WMT17 2nd Winner (Martins et al., 2017)	0.65	0.63
NeuralTQE	0.41	**0.45**

Table 7.6 Predicting MT HTER scores

	DE-EN		EN-DE			
	SMT		NMT		SMT	
	sentences	words	sentences	words	sentences	words
training	26032	493010/509456	13442	234725/255610	26299	442074/466026
development	1000	18817/19434	1000	17669/19224	1000	16565/17462
test	2000	34793/36163	2000	33643/35577		

Table 7.7 Staistics of the German-English and English-German datasets

Post-Editing Time Prediction

The WMT 2018 quality estimation shared task [5] also provides additional labels collected during post-editing:

- post-editing time in seconds
- number of keys pressed for ten types of keys

To verify the workability of the proposed model, I train models for estimating post-editing time and the summation of keystrokes needed in a revision for each MT sentence. I choose the DE-EN and EN-DE and English-Czech (EN-CS) data[6] released by the organiser (Specia et al., 2018). Note that the EN-DE data contains both SMT and NMT data and I report post-editing time and keystrokes for two types of MT data in this language direction. Table 7.7 lists the data statistics. Table 7.8 details the results of models on the development datasets[7].

According to the results in Table 7.6 and Table 7.8, we can see NeuralTQE models perform rather stable across different tasks, having achieved the same level of performance ($r \geqslant 0.3$) on both HTQE data and MTQE data on task-specific QE. In particular, models for estimating post-editing time and keystrokes for revision in the

[5] http://www.statmt.org/wmt18/quality-estimation-task.html
[6] We kept experiencing error dealing with the Latvian data and finally gave up.
[7] We do not have the gold labels for the test data at the moment of writing.

	DE-EN	**EN-DE**		**EN-CS**
Target	SMT	NMT	SMT	SMT
PE Time	0.41	0.44	0.35	0.55
Key Strokes	0.33	0.40	0.42	0.73

Table 7.8 Predicting MT post-editing Time and Keystrokes

English-Czech direction achieve the highest correlation with reference annotations, using no more information than word embeddings.

7.8 Summary

I have introduced a neural model for human translation quality estimation, proposing a weighted cross attention mechanism to adaptively detect the essential parts in source-target sentence pairs for quality estimation. Without feature engineering, results show that the neural model can outperform conventional feature-based baselines. This research can be regarded as a step towards fully automated reference-free translation quality evaluation using neural models.

This chapter makes several noteworthy contributions to current translation quality estimation studies. A stacked neural framework leveraging both CNN and RNN networks with an attention mechanism is proposed to model translation pairs. Besides, this work treats sentence-level HTQE as a learning task of simulating fine-grained direct assessments of human annotators, providing insight into human translation quality from different aspects and thus possibly more useful feedback to trainee translators.

I have also shown the proposed feature framework can be leveraged to build models for the sentence-level HTQE. In comparison to feature-based models, the neural framework has achieved impressive performance to estimate fine-grained quality scores for sentence-level trainee translations with a much simpler structure design and has demonstrated its application in the MTQE task to predict post-editing efforts.

I identify that the word representation based neural model favours literal translations, and it suffers insensitivity to meaning changes resulting from lexical changes.

Chapter 8

Conclusion and Further Research

HTQE, as a nascent, interdisciplinary subfield of Translation Studies, calls for more in-depth methodological and theoretical innovation. The issue of HTQE has only recently drawn the attention of a few translation scholars and language researchers (Yuan, 2016), and most quantitative analysis of human translations quality are not designed with a 'reference-free estimation' in mind. Reference-free fine-grained HTQE has a significant potential for translator education, translation quality control and qualification accreditation, fitting in the scenario where evaluating human translations can be automatised for different tasks, e.g. pedagogical, self-learning, quality control and scoring for large-scale certification examinations.

In this thesis, I have attempted to address the issue of HTQE from two different approaches: feature-based and deep learning-based. Throughout the work, I have been working on two challenges: translation representation and learning effectiveness. The challenge of translation representation is related to the fact that HTs are often a textual-linguistic product to a large extent different from MTs. The first goal has been to *exploit methods and algorithms to develop novel representations that are useful for building effective quality estimation models for HTs*. In response, I employ a selection of features with high predictive power from previous research in MTQE, while engineering extensively new features that aim to capture multiple aspects of translation quality (e.g. content, style, terminology, discourse information). The second goal is associated with the effective learning, i.e. *to compare effective machine learning QE models for fine-grained quality prediction*. Rather than automatic metric scores and post-editing effort metric (HTER), I tried to predict a hierarchy of 7 quality scores for translated documents and sentences that have been assigned by human annotators. I investigated the effectiveness of conventional feature-based statistical learning algorithms and deep neural networks. In this chapter, I recap the main findings of the thesis, discuss its contribution, foreground its implications, and highlight potential issues to address in the future.

8.1 Conclusions

The main **conclusions** of this work are:

- for **RQ1** and **RQ2**
  - Monolingual, bilingual and language modelling features spanning shallow, lexical, syntactic, discourse aware features and terminology features, on top of a small proportion of MTQE-inspired features, comprise the framework of features that has been designed for interpretability (association of the features with translation quality) and simplicity (ease of computation, avoiding computational complexity and reliance on too many external resources).
  - Using language-independent statistical features, I identify terminology from monolingual texts via a supervised classifier. N-grams from both STs and TTs can be classified automatically as terms or non-terms. Trained with cross-domain and cross-language data, the term classifier based on the proposed feature set of statistics, with mediocre precision but satisfactory recall, could help identify potential terms in the target texts of technical domains, where normalised frequencies of such 'terms' are indicative of translation quality in different aspects, showing above-weak to moderate correlation with multiple quality scores.
- for **RQ3** and**RQ4**
  - At the document level, with the proposed feature set, quality estimation can be achieved through regression learning. Fine-grained quality scores have been predicted with high correlation with human judgements ($\mathrm{r} > 0.7$) via training on human annotated examples. Overall, feature selected using variable ranking have demonstrated the possibility of outperforming the strong baseline for estimating different quality components in different configurations of learning settings. When applied to a small set of human annotated MT data, the model based on the selected k-best features performs most effectively ($\mathrm{r} > 0.49$), while two other models (**Full** and **TopNCorr**) outperform the baseline **QuEst** in most aspects except predicting the total score.
  - The proposed feature framework is suitable for both document-level and sentence-level HTQE, giving the best estimators with the selected features on the data. The proposed neural model ranked the second best among four other feature-based models for sentence-level HTQE. When the neural framework is applied to estimating HTER scores in MTQE, it has achieved comparable performance to the strong baseline

in the WMT17 quality estimation shared task, showing strong potential for other task-oriented MTQE as well, e.g. predicting post-editing time, keystrokes, with only word embeddings as input.

– Principal component analysis of the distribution of translation errors reveals that human translation quality issues are more content-related, which are largely due to translator's decision making and incapability of switching between two languages. In contrast, MT translation errors are more related to target language quality. MT errors occur more in creative texts (e.g. prose), and HT errors are more pronounced in Science texts. Some categories of features, e.g. shallow features, bilingual and distance metrics highly correlate with all quality scores. In particular, target-side features are more prominent than source-side features.

The main **contributions** of this thesis are:

- **Extension of translation quality features.** I have developed a large feature set for HTQE, of which the majority are newly engineered. Experiments show these features can be used for estimating various quality components of a fine-grained evaluation scheme (ATA rubric 2011). Though the features are designed with document-level comparison in mind, most of these features are applicable to the sentence representation.

- **Supervised learning for cross-lingual term extraction.** I have proposed a supervised learning method for classifying bilingual n-grams into terms and non-terms for the purpose of quality estimation. Different from the statistical and linguistic approach, I exploit statistical features that are independent of languages to train classification models and apply them to cross-language and cross-domain texts. On the same dataset (GENIA), the F-score of the model increases from 0.72 to 0.82. Experiments show the number of terms identified moderately correlates with human annotated quality scores.

- **Corpus of Translations with quality annotation.** I have built a corpus of English-Chinese trainee translations with fine-grained quality scores. I sampled 457 documents and 3529 sentence pairs of trainee translations from a published parallel corpus, and have them manually annotated with quality scores according to the ATA rubric. I have also annotated a set of machine and human translations (42 documents each) using the adapted MQM-DQF synthetic framework. On its basis, I ran the principal component analysis to reveal the interaction of translation text types and errors. The results are useful to describe and explain the translation errors in HTs and MTs. At the same time, findings from such data-driven analysis may lend insights into future feature engineering for more robust models.

- **Investigation of individual feature contribution to different translation quality aspects.** I have investigated the contribution of individual features to specific quality scores. I used the pair-wise correlation analysis to rank the linear dependency between each feature and each quality label in the quality annotation scheme. I further investigated their contribution using the 'Kbest' feature selection method in association with grid search. This comprehensive study allows patterns of feature-quality association to emerge. The detailed correlations between them can point a direction to build predictive quality estimation models, as illustrated in Chapter 7. Large-scale analysis of the contribution of textual-linguistic features to multi-level quality scores has not been done before this study.

- **HTQE at different granularities.** I have investigated HTQE at the document- and sentence-level. I used the XGBoost algorithm with grid search parameter optimisation to predict fine-grained quality scores for HTs, and even MTs at the document level, having achieved a 0.62-0.76 correlation with human judgements at the document-level, and 0.34 -0.55 at the sentence-level.

- **Stacked neural model for sentence-level HTQE.** I have proposed a hierarchical neural model with a customised attention mechanism to capture sentence level equivalence. The proposed neural model proves effective for reference-free quality estimation at the sentence level. For HTQE, the proposed method has obtained a marginal increase of 4-8% in terms of the correlations with human judgements. For MTQE, the neural model is comparable to the baseline but falls behind the winning systems.

8.2 Further Work

However, I shall acknowledge the **limitations** of my efforts in this thesis, and propose some **directions for future work**.

- **Extending the branches of current translation quality assessment in Translation Studies.**
  In Holmes's 'map' of Translation Studies, quality assessment is only a branch of Applied Translation Studies under the umbrella of translation criticism (Toury, 2012). This phenomenon conflicts with the truth that artificial intelligence has become an inseparable part of our life and research, with almost all quality assessment related studies centring around manual appreciation and evaluation. That the evaluation process of translation quality can be computerised has been long ignored in the world of Translation Studies. This study showcases an undertaking in which that HTs are represented in numeric vectors through carefully engineered features and/or via unsupervised representation learning method so as to automate their evaluation at different granularities.

Automatising the evaluation process for specific scenarios where a reliable quality system can provide fine-grained feedback to end users is what I attempt to achieve. However, more systemic and in-depth work and condense conceptualisation are warranted. In the future, interdisciplinary efforts combining the domain knowledge of Translation Studies and statistical learning could be employed to enrich and deepen this subfield.

- **Integrating ATR with HTQE more elegantly.**
  The attempt to automatise the term extraction for TQE could be further extended for domain-specific TQE in the future. I will inspect more refined ways of integrating ATR results with TQE task, exploring how the term identification information can be maximally utilised. For instance, I plan to treat the terminology extraction as a neural sequence labelling task. Without relying on handcrafted features and task-specific resources, I will examine the plausibility of neural models for terminology extraction. Many neural models have obtained state-of-the-art performance on other sequence labelling tasks (Lample et al., 2016; Ma and Hovy, 2016–2018). Therefore, I can train a neural model on the monolingual in-domain data by utilising character- (Santos and Zadrozny, 2014) and word-level CNN (Strubell et al., 2017) and LSTM structures to represent global sequence information, in addition to a Conditional Random Field (CRF) (Lafferty et al., 2001) layer to capture dependencies between neighbouring labels.

- **Promoting multi-purpose HTQE applications.**
  At the document level, the feature-based models have achieved high correlations with human judgements (up to 0.7 in terms of Pearson). This suggests that these working models can be utilised in scenarios where fast, cheap and consistent human-like evaluation is needed. For instance, incorporating these models into computer-aided translation learning environment to offer quality feedback to trainee translators could be promising. With these models, a translation could be selected by the criterion of certain quality aspect, e.g. low **UT** scores, average **TM** scores, for post-editing or diagnosis analysis. Perhaps, the most pertinent and valuable application of these models to this study is to exploit these models in large-scale translation exams, e.g. university-wide, nation-wide and even world-wide certification exams, to grade translation work at different granularities, providing assistance to human evaluators just like the e-rater® automated writing evaluation engine to ETS essay scoring (Burstein, 2003).

- **Applying the feature framework to other NLP tasks.**
  The large set of features can be potentially used for other text analytical tasks, such as stylometry comparison, document classification and language identification. For instance, I can utilise the proposed feature-based method to

identify translated and non-translated texts and see how they perform against other methods (Ilisei et al., 2010; Rabinovich and Wintner, 2015). In particular, I may apply the feature framework and the neural model to filter low-quality translations in classification tasks.

- **Optimising parameters.**
  Even though document-level quality estimation models built with the proposed features together have manifested superiority in estimating all specified quality scores at both the document-level and the sentence-level, this better performance, however, is not best optimised. Though I have carried out grid search to fine tune the models and implement feature selection for each quality component, I specified a very narrow range of parameters to cut off running time. In the future, I need to experiment with more optimisation for each submodel.

- **Improving the neural model.**
  The proposed neural model with attention mechanism achieves reasonably good correlation with human judgements at the sentence level for direct human assessment quality scores (e.g. ATA scores). It also has achieved fairly reasonable performance for indirect quality measures (e.g. HTER scores, keystrokes) on the MT data. The stacked neural framework leverages the strengths of two neural architectures, using only continuous dense vectors learned from monolingual corpora. This fact leaves space for future work to

  – **integrate discrete feature layers.**
    Manual features have been investigated for most NLP tasks over the past decades and cover the most useful indicators for solving the problem. For QE, such information can be complementary to features automatically induced from neural networks. Therefore, combining discrete features with neural representations can potentially lead to improved model performance. For example, it is worth incorporating alignment information into the current model.

  – **adjust the structure of the encoding layers.**
    I have employed BiLSTM on top of a CNN convolutional layer with the attention to encode ST-TT pairs. However, this structure can be adjusted in many ways. For instance, I can switch the order of CNN and LSTM architecture for different effects. Alternatively, I may try other improved neural models such as soft patterns, a neural version of a weighted finite state automation (Mohri, 1996) that extends the one-layer CNN and lies in between RNNs and CNNs (Schwartz et al., 2018), and the LSTM-state (Zhang et al., 2018), a parallel state LSTM recurrently exchanging local and global information simultaneously rather than incrementally, that have shown stronger representation power for sentences in text classification tasks.

– **implement different attention strategies.**
For easy computation, I adopt a linear transformation of the encoded sentence representation and corresponding hidden status of words in the counterpart sentence. Despite that there are signs this attention mechanism has improved the neural model in comparison to the one without attention, other attention implementation methods could be potentially advantageous. In particular, multi-head self-attention (Vaswani et al., 2017a), bilinear dot-product method (Vaswani et al., 2017b), attention weights control by coverage constraint (Luong et al., 2015b), and attention weights sharpening method (Chorowski et al., 2015), e.g. by adding temperature parameter β to the softmax function, or masking low weight, have been recently proposed for NMT and speech recognition. The effectiveness of these methods on modelling sentence pairs in terms of translation quality needs further investigation.

– **train task-specific embeddings.**
In comparison to the winning systems (Kim et al., 2017; Martins et al., 2017) for WMT17 QE shared task, the proposed neural model needs to be improved. I notice that Martins et al. (2017) have achieved the state-of-the-art performance by training a neural word prediction model from parallel corpora and combining it with the neural quality estimation model. Their stacked model contains the quality information of word translations. In contrast, in this thesis the neural model uses only word embeddings trained from monolingual corpora. This is problematic for TQE as the word-level quality information has been ignored. Recently, task-specific word embeddings have been considered important for various NLP tasks (Tang et al., 2014; Gouws and Søgaard, 2015), and new algorithms (Artetxe et al., 2018; Athiwaratkun et al., 2018) to train embeddings more effectively for a range of tasks have been advanced. I intend to address this issue by learning quality-specific word embeddings that integrate the quality information into the loss function of the chosen neural networks, seeing if new methods of sub-word, word and sentence representation can improve the model accuracy.

– **compare with other neural models.**
Due to the nature of this research is to investigate the effectiveness of the representation means of translations, it is thus not prioritised to explore efficacy of different deep learning architectures. For this reason, I compared only the feature-based model with the proposed neural model. No other neural models are compared to the proposed method. It would be fair to compare a single CNN, LSTM or BiLSTM model with the stacked model in this thesis.

- **Building high-quality HT data for TQE shared task.**
  The datasets presented in this study are still relatively small. Both translated documents and sentences need to be expanded regarding the number of samples and language directions. For instance, due to the small number of document-level translations, I could not examine the effectiveness of the proposed neural model for document-level QE. In the future, I consider crowdsourcing the quality annotation for a larger number of annotated data of multiple language directions. These data can be employed to organise fine-grained QE shared tasks similar to current WMT QE shared tasks.

- **Implementing HTQE for multiple language pairs.**
  Due to the lack of knowledge of other languages and time constraint, I could not investigate the performance of feature-based models for translations of language pairs other than English-Chinese. It would be worthwhile to explore the effectiveness of the proposed feature set for MTQE of different language pairs at different granularities.

References

Abdelsalam, Amal, Bojar, Ondřej, and El-Beltagy, Samhaa (2016). “Bilingual embeddings and word alignments for translation quality estimation”. In: *Proceedings of the First Conference on Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Rajan Chatterjee, Christian Federmann, Liane Guillou, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Aurelie Neveol, Mariana Neves, Pavel Pecina, Martin Popel, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Lucia Specia, Karin Verspoor, Joerg Tiedemann, and Marco Turchi. Berlin, Germany: Association for Computational Linguistics, Aug. 11–12, 2016, pp. 764–771.

Abdi, Hervé and Williams, Lynne J. (2010). “Principal component analysis”. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), pp. 433–459.

Agarwal, Abhaya and Lavie, Alon (2008). “Meteor, M-BLEU and M-TER: Evaluation metrics for high-correlation with human rankings of machine translation output”. In: *Proceedings of the Third Workshop on Statistical Machine Translation*. Ed. by Chris Callison-Burch, Philipp Koehn, Christof Monz, Josh Schroeder, and Cameron Shaw Fordyce. Columbus, Ohio: Association for Computational Linguistics, June 19, 2008, pp. 115–118.

Ahmad, Khurshid, Gillam, Lee, and Tostevin, Lena (1999). “Weirdness indexing for logical document extrapolation and retrieval”. In: *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*. Ed. by Ellen Voorhees and Donna Harman. Gaithersburg, Maryland, USA: National Institute of Standards and Technology, Nov. 16–19, 1999, pp. 717–724.

Ahrenberg, Lars (2017). “Comparing machine translation and human translation: A case study”. In: ed. by Irina Temnikova, Constantin Orasan, Gloria Corpas Pastor, and Stephan Vogel. Varna, Bulgaria: Association for Computational Linguistics, Shoumen, Bulgaria, Sept. 7, 2017, pp. 21–28.

Albrecht, Joshua and Hwa, Rebecca (2007a). “A re-examination of machine learning approaches for sentence-level MT evaluation”. In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Ed. by Antal van den Bosch and Annie Zaenen. Prague, Czech Republic: Association for Computational Linguistics, June 23–30, 2007, pp. 880–887.

Albrecht, Joshua and Hwa, Rebecca (2007b). “Regression for sentence-level MT evaluation with pseudo references”. In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Ed. by Antal van den Bosch and Annie Zaenen. Prague, Czech Republic: Association for Computational Linguistics, June 23–30, 2007, pp. 296–303.

Almaghout, Hala and Specia, Lucia (2013). “A CCG-based quality estimation metric for statistical machine translation”. In: *Proceedings of Machine Translation Summit XIV*. Ed. by Andy Way, Khalil Sima’an, Mikel L. Forcada, Daniel Grasmick, and Heidi Depaetere. Nice, France: International Association for Machine Translation, Sept. 2–6, 2013, pp. 223–230.

Ananiadou, Sophia (1994). “A methodology for automatic term recognition”. In: *Proceedings of the 15th Conference on Computational Linguistics-Volume 2*. Ed. by

Makoto Nagao, Yorick Wilks, Shigeru Sato, Hozumi Tanaka, Kohei Habara, and Masahide Yoneyama. Kytoto, Japan: Association for Computational Linguistics, Aug. 5–9, 1994, pp. 1034–1038.

Angelelli, Claudia and Jacobson, Holly (2009a). "Testing and Assessment in Translation and Interpreting Studies:A call for dialogue between research and practice". In: ed. by Claudia Angelelli and Holly Jacobson. Amsterdam / Philadelphia: John Benjamins Publishing Company. Chap. introduction:Testing and assessment in translation and interpreting studies: A call for dialogue between research and practice, pp. 1–10.

Angelelli, Claudia and Jacobson, Holly (2009b). "Testing and Assessment in Translation and Interpreting Studies:A call for dialogue between research and practice". In: ed. by Claudia Angelelli and Holly Jacobson. Amsterdam / Philadelphia: John Benjamins Publishing Company. Chap. Using a rubric to assess translation ability: Defining the construct, pp. 14–49.

Artetxe, Mikel, Labaka, Gorka, and Agirre, Eneko (2016). "Learning principled bilingual mappings of word embeddings while preserving monolingual invariance". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 1–6, 2016, pp. 2289–2294.

Artetxe, Mikel, Labaka, Gorka, and Agirre, Eneko (2017). "Learning bilingual word embeddings with (almost) no bilingual data". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Regina Barzilay and Min-Yen Kan. Vancouver, Canada: Association for Computational Linguistics, July 30–Aug. 4, 2017, pp. 451–462.

Artetxe, Mikel, Labaka, Gorka, and Agirre, Eneko (2018). "A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Iryna Gurevych and Yusuke Miyao. Melbourne, Australia: Association for Computational Linguistics, July 15–20, 2018, pp. 789–798.

ATA (2010). *Codes of ethics for translators and interpreters*. Ed. by American Translators Association. [Online] Available from http://www.monicamartinez.es/ethicscode.pdf. [Accessed on 25 July 2017]. Oct. 2010.

ATA (2011). *ATA certification program rubric for grading*. Ed. by American Translators Association. Atanet.org. [Online] Available from http://www.atanet.org/certification/aboutexams_rubic.pdf. [Accessed on 8 May, 2017]. Oct. 2011.

ATA (2017). *Explanation of error categories*. Ed. by American Translators Association. [Online] Available from https://www.atanet.org/certification/aboutexams_error.php. [accessed on 1 April 2018]. Oct. 2017.

Athiwaratkun, Ben, Wilson, Andrew, and Anandkumar, Anima (2018). "Probabilistic FastText for Multi-Sense Word Embeddings". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Iryna Gurevych and Yusuke Miyao. Melbourne, Australia: Association for Computational Linguistics, July 15–20, 2018, pp. 1–11.

Avramidis, Eleftherios (2012a). "Comparative quality estimation: Automatic sentence-level ranking of multiple machine translation outputs". In: *Proceedings of 24th International Conference on Computational Linguistics*. Ed. by Martin Kay and Christian Boitet. Mumbai, India: The COLING 2012 Organizing Committee, Dec. 8–15, 2012, pp. 115–132.

Avramidis, Eleftherios (2012b). "Quality estimation for machine translation output using linguistic analysis and decoding Features". In: *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Ed. by Chris Callison-Burch, Philipp

Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. WMT '12. Montréal, Canada: Association for Computational Linguistics, June 7–8, 2012, pp. 84–90.

Avramidis, Eleftherios (2014). "Efforts on machine learning over human-mediated translation edit rate". In: *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Matouš Macháček, Christof Monz, Pavel Pecina, Matt Post, Hervé Saint-Amand, Radu Soricut, and Lucia Specia. Baltimore, Maryland, USA: Association for Computational Linguistics, June 26–27, 2014, pp. 302–306.

Avramidis, Eleftherios and Popovic, Maja (2013a). "Machine learning methods for comparative and time-oriented quality estimation of machine translation output". In: *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Chris Callison-Burch, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Hervé Saint-Amand, Radu Soricut, and Lucia Specia. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 8–9, 2013, pp. 329–336.

Avramidis, Eleftherios and Popovic, Maja (2013b). "Selecting feature sets for comparative and time-oriented quality estimation of machine translation output". In: *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Chris Callison-Burch, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Hervé Saint-Amand, Radu Soricut, and Lucia Specia. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 8–9, 2013, pp. 329–336.

Azzopardi, Leif, Girolami, Mark, and Risjbergen, Keith van (2003). "Investigating the relationship between language model perplexity and IR precision-recall measures". In: *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*. Ed. by Charles Clarke, Gordon Cormack, Jamie Callan, David Hawking, and Alan Smeaton. SIGIR '03. Toronto, Canada: Association for Computing Machinery, July 28–Aug. 1, 2003, pp. 369–370.

Baayen, R.H., Davidson, D.J., and Bates, D.M. (2008). "Mixed-effects modeling with crossed random effects for subjects and items". *Journal of Memory and Language*, 59(4). Special Issue: Emerging Data Analysis, pp. 390–412.

Babych, Bogdan (2014). "Automated MT evaluation metrics and their limitations". *Tradumàtica*, (12), pp. 464–470.

Babych, Bogdan and Hartley, Anthony (2004). "Extending the BLEU MT evaluation method with frequency weightings". In: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Ed. by Donia Scott, Walter Daelemans, and Marilyn Walk er. ACL '04. Barcelona, Spain: Association for Computational Linguistics, July 21–26, 2004, pp. 621–628.

Bach, Nguyen, Huang, Fei, and Al-Onaizan, Yaser (2011). "Goodness: A method for measuring machine translation confidence". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Techologies*. Ed. by Dekang Lin, Yuji Matsumoto, and Rada Mihalcea. Portland, Oregon, USA: Association for Computational Linguistics, June 19–24, 2011, pp. 211–219.

Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua (2015). "Neural machine translation by jointly learning to align and translate". In: *Proceedings of the 3rd International Confererence on Learning Representations, ICLR 2015*. Ed. by Brian Kingsbury, Samy Bengio, ando de Freitas, and Hugo Larochelle. Sandiego, USA, May 7–9, 2015.

Ballard, Michel (2010). “Unit of translation”. In: *Handbook of Translation Studies*. Ed. by Yves Gambier and Luc Van Doorslaer. Vol. 1. Amsterdam/Philadelphia: John Benjamins Publishing, pp. 437–440.

Banerjee, Satanjeev and Lavie, Alon (2005). “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments”. In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ed. by Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss. Ann Arbor, Michigan: Association for Computational Linguistics, June 29, 2005, pp. 65–72.

Beck, Daniel, Shah, Kashif, Cohn, Trevor, and Specia, Lucia (2013). “SHEF-Lite: when less is more for translation quality estimation”. In: *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Chris Callison-Burch, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Hervé Saint-Amand, Radu Soricut, and Lucia Specia. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 8–9, 2013, pp. 337–342.

Beck, Daniel, Shah, Kashif, and Specia, Lucia (2014). “SHEF-Lite 2.0: sparse multi-task gaussian processes for translation quality estimation”. In: *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Matouš Macháček, Christof Monz, Pavel Pecina, Matt Post, Hervé Saint-Amand, Radu Soricut, and Lucia Specia. Baltimore, Maryland, USA: Association for Computational Linguistics, June 26–27, 2014, pp. 307–312.

Beck, Daniel, Vlachos, Andreas, Paetzold, Gustavo, and Specia, Lucia (2016). “SHEF-MIME: Word-level quality estimation using imitation learning”. In: *Proceedings of the First Conference on Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Rajan Chatterjee, Christian Federmann, Liane Guillou, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Aurelie Neveol, Mariana Neves, Pavel Pecina, Martin Popel, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Lucia Specia, Karin Verspoor, Joerg Tiedemann, and Marco Turchi. Berlin, Germany: Association for Computational Linguistics, Aug. 11–12, 2016, pp. 772–776.

Ben, Guosheng, Xiong, Deyi, Teng, Zhiyang, Lü, Yajuan, and Liu, Qun (2013). “Bilingual lexical cohesion trigger model for document-level machine translation”. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Pascale Fung and Masimo Poesio. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 4–9, 2013, pp. 382–386.

Bengio, Yoshua, Courville, Aaron, and Vincent, Pascal (2013). “Representation learning: A review and new perspectives”. *IEEE transactions on pattern analysis and machine intelligence*, 35(8), pp. 1798–1828.

Bengio, Yoshua, Ducharme, Réjean, Vincent, Pascal, and Jauvin, Christian (2003). “A neural probabilistic language model”. *Journal of Machine Learning Research*, 3(Feb), pp. 1137–1155.

Bergstra, James and Bengio, Yoshua (2012). “Random search for hyper-parameter optimization”. *Journal of Machine Learning Research*, 13(Feb), pp. 281–305.

Bicici, Ergun (2013). “Referential translation machines for quality estimation”. In: *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Chris Callison-Burch, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Hervé Saint-Amand, Radu Soricut, and Lucia Specia. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 8–9, 2013, pp. 343–351.

Bicici, Ergun (2016). “Referential translation machines for predicting translation performance”. In: *Proceedings of the First Conference on Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Rajan Chatterjee, Christian Federmann, Liane Guillou, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Aurelie Neveol, Mariana Neves, Pavel Pecina, Martin Popel, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Lucia Specia, Karin Verspoor, Joerg Tiedemann, and Marco Turchi. Berlin, Germany: Association for Computational Linguistics, Aug. 11–12, 2016, pp. 777–781.

Bjerva, Johannes, Bos, Johan, Goot, Rob van der, and Nissim, Malvina (2014). “The meaning factory: Formal semantics for recognizing textual entailment and determining semantic similarity”. In: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Ed. by Preslav Nakov and Torsten Zesch. Dublin, Ireland: Association for Computational Linguistics and Dublin City University, Aug. 23–24, 2014, pp. 642–646.

Blancafort, Helena, Daille, Béatrice, Gornostay, Tatiana, Heid, Ulrich, Méchoulam, Claude, and Sharoff, Serge (2010). “TTC: Terminology extraction, translation tools and comparable corpora”. In: *Proceedings of the XIV Euralex International Congress*. Ed. by Anne Dykstra and Tanneke Schoonheim. Leeuwarden/Ljouwert, Netherlands: Fryske Akademy, July 6–10, 2010, pp. 263–268.

Blatz, John, Fitzgerald, Erin, Foster, George, Gandrabur, Simona, Goutte, Cyril, Kulesza, Alex, Sanchis, Alberto, and Ueffing, Nicola (2004). “Confidence estimation for machine translation”. In: *Proceedings of the 20th International Conference on Computational Linguistics*. COLING ’04. Geneva, Switzerland: Association for Computational Linguistics, Aug. 23–27, 2004, pp. 315–321.

Blei, David M, Ng, Andrew Y, and Jordan, Michael I (2003). “Latent dirichlet allocation”. *Journal of Machine Learning Research*, 3(Jan), pp. 993–1022.

Bojanowski, Piotr, Grave, Edouard, Joulin, Armand, and Mikolov, Tomas (2016). “Enriching word vectors with subword information”. *CoRR*, abs/1607.04606. [Online] Available from http://arxiv.org/abs/1607.04606. [Accessed on 23 April 2017].

Bojar, Ondřej (2011). “Analyzing Error Types in English-Czech Machine Translation”. *Prague Bulletin of Mathematical Linguistics* (95 Apr. 2011), pp. 63–76.

Bojar, Ondřej, Buck, Christian, Callison-Burch, Chris, Federmann, Christian, Haddow, Barry, Koehn, Philipp, Monz, Christof, Post, Matt, Soricut, Radu, and Specia, Lucia (2013). “Findings of the 2013 workshop on statistical machine translation”. In: *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Chris Callison-Burch, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Hervé Saint-Amand, Radu Soricut, and Lucia Specia. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 8–9, 2013, pp. 1–44.

Bojar, Ondřej, Buck, Christian, Federmann, Christian, Haddow, Barry, Koehn, Philipp, Leveling, Johannes, Monz, Christof, Pecina, Pavel, Post, Matt, Saint-Amand, Herve, Soricut, Radu, Specia, Lucia, and Tamchyna, Aleš (2014). “Findings of the 2014 workshop on statistical machine translation”. In: *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Matouš Macháček, Christof Monz, Pavel Pecina, Matt Post, Hervé Saint-Amand, Radu Soricut, and Lucia Specia. Baltimore, Maryland, USA: Association for Computational Linguistics, June 26–27, 2014, pp. 12–58.

Bojar, Ondřej, Chatterjee, Rajen, Federmann, Christian, Graham, Yvette, Haddow, Barry, Huang, Shujian, Huck, Matthias, Koehn, Philipp, Liu, Qun, Logacheva,

Varvara, Monz, Christof, Negri, Matteo, Post, Matt, Rubino, Raphael, Specia, Lucia, and Turchi, Marco (2017). “Findings of the 2017 conference on machine translation (WMT17)”. In: *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 7–8, 2017, pp. 169–214.

Bojar, Ondřej, Chatterjee, Rajen, Federmann, Christian, Graham, Yvette, Haddow, Barry, Huck, Matthias, Jimeno Yepes, Antonio, Koehn, Philipp, Logacheva, Varvara, Monz, Christof, Negri, Matteo, Neveol, Aurelie, Neves, Mariana, Popel, Martin, Post, Matt, Rubino, Raphael, Scarton, Carolina, Specia, Lucia, Turchi, Marco, Verspoor, Karin, and Zampieri, Marcos (2016a). “Findings of the 2016 conference on machine translation”. In: *Proceedings of the First Conference on Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Rajan Chatterjee, Christian Federmann, Liane Guillou, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Aurelie Neveol, Mariana Neves, Pavel Pecina, Martin Popel, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Lucia Specia, Karin Verspoor, Joerg Tiedemann, and Marco Turchi. Berlin, Germany: Association for Computational Linguistics, Aug. 11–12, 2016, pp. 131–198.

Bojar, Ondřej, Chatterjee, Rajen, Federmann, Christian, Haddow, Barry, Huck, Matthias, Hokamp, Chris, Koehn, Philipp, Logacheva, Varvara, Monz, Christof, Negri, Matteo, Post, Matt, Scarton, Carolina, Specia, Lucia, and Turchi, Marco (2015). “Findings of the 2015 workshop on statistical machine translation”. In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, Pavel Pecina, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Hervé Saint-Amand, Carolina Scarton, Lucia Specia, and Marco Turchi. Lisbon, Portugal: Association for Computational Linguistics, Sept. 17–18, 2015, pp. 1–46.

Bojar, Ondřej, Graham, Yvette, Kamran, Amir, and Stanojević, Miloš (2016b). “Results of the WMT16 metrics shared task”. In: *Proceedings of the First Conference on Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Rajan Chatterjee, Christian Federmann, Liane Guillou, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Aurelie Neveol, Mariana Neves, Pavel Pecina, Martin Popel, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Lucia Specia, Karin Verspoor, Joerg Tiedemann, and Marco Turchi. Berlin, Germany: Association for Computational Linguistics, Aug. 11–12, 2016, pp. 199–231.

Bonial, Claire, Hwang, Jena, Bonn, Julia, Conger, Kathryn, Babko-Malaya, Olga, and Palmer, Martha (2012). *English propbank annotation guidelines*. Center for Computational Language and Education Research Institute of Cognitive Science University of Colorado. [Online] Availble from http://verbs.colorado.edu/propbank/EPB-Annotation-Guidelines.pdf. [Accessed on 25 May 2017]. Boulder, Colorado, Nov. 24, 2012.

Bourigault, Didier, Gonzalez-Mullier, Isabelle, and Gros, Cécile (1996). “LEXTER, a natural language processing tool for terminology extraction”. In: *Proceedings of the 7th EURALEX International Congress*. Ed. by Martin Gellerstam, Jerker Järborg, Sven-Göran Malmgren, Kerstin Norén, Lena Rogström, and Catarina Röjder Papmehl. Göteborg, Sweden, Aug. 13–18, 1996, pp. 771–779.

Bowman, Samuel R., Angeli, Gabor, Potts, Christopher, and Manning, Christopher D. (2015). “A large annotated corpus for learning natural language inference”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Ed. by Lluís Màrquez, Chris Callison-Burch, and Jian Su. Lisbon, Portugal: Association for Computational Linguistics, Sept. 17–21, 2015, pp. 632–642.

Breiman, Leo (2001). “Random Forests”. *Machine Learning*, 45(1) (Oct. 2001), pp. 5–32.

Brislin, Richard W. (1995). “Back-translation: A tool for cross-cultural research”. In: *An encyclopaedia of translation: Chinese-English, English-Chinese*. Ed. by Sin-wai Chan and David E. Pollard. Hong Kong: The Chinese University Press, pp. 22–40.

Brown, Peter F, Pietra, Vincent J Della, Pietra, Stephen A Della, and Mercer, Robert L (1993). “The mathematics of statistical machine translation: Parameter estimation”. *Computational linguistics*, 19(2), pp. 263–311.

Brunette, Louise (2000). “Towards a terminology for translation quality assessment”. *The Translator*, 6(2), pp. 169–182.

Buck, Christian (2012). “Black box features for the wmt 2012 quality estimation shared task”. In: *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Ed. by Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Montréal, Canada: Association for Computational Linguistics, June 7–8, 2012, pp. 88–92.

Buduma, Nikhil and Lacascio, Nicholas (2017). *Fundamentals of deep learning: Designing next-generation machine intelligence algorithms*. Sebastopol, California: O’Reilly Media, Inc., June 2017.

Burstein, Jill C. (2003). “The E-rater® scoring engine: Automated essay scoring with natural language processing”. In: *Automated Essay Scoring: A Cross-disciplinary Perspective*. Ed. by Mark D. Shermis and Jill C. Burstein. New York: Routledge, pp. 107–116.

Cabré, M. Teresa (2010). “Terminology and translation”. In: *Handbook of Translation Studies*. Ed. by Yves Gambier and Luc Van Doorslaer. Vol. 1. Amsterdam/Philadelphia: John Benjamins Publishing, pp. 356–366.

Callison-Burch, Chris, Koehn, Philipp, Monz, Christof, Peterson, Kay, and Zaidan, Omar (2010). “Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR”. In: Uppsala, Sweden: Association for Computational Linguistics, July 15–16, 2010.

Callison-Burch, Chris, Koehn, Philipp, Monz, Christof, Post, Matt, Soricut, Radu, and Specia, Lucia (2012). “Findings of the 2012 workshop on statistical machine translation”. In: *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Ed. by Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Montréal, Canada: Association for Computational Linguistics, June 7–8, 2012, pp. 10–51.

Callison-Burch, Chris, Osborne, Miles, and Koehn, Philipp (2006). “Re-evaluation the role of Bleu in machine translation research”. In: *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*. Ed. by Frank Keller and Gabor Proszeky. Trento, Italy: Association for Computational Linguistics, Apr. 5–6, 2006, pp. 249–256.

Camargo de Souza, José Guilherme, Buck, Christian, Turchi, Marco, and Negri, Matteo (2013). “FBK-UEdin participation to the WMT13 quality estimation shared task”. In: *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Chris Callison-Burch, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Hervé Saint-Amand, Radu Soricut, and Lucia Specia. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 8–9, 2013, pp. 352–358.

Camargo de Souza, José Guilherme, González-Rubio, Jesús, Buck, Christian, Turchi, Marco, and Negri, Matteo (2014). “FBK-UPV-UEdin participation in the

WMT14 quality estimation shared-task". In: *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Matouš Macháček, Christof Monz, Pavel Pecina, Matt Post, Hervé Saint-Amand, Radu Soricut, and Lucia Specia. Baltimore, Maryland, USA: Association for Computational Linguistics, June 26–27, 2014, pp. 322–328.

Carreras, Xavier and Màrquez, Lluís (2005). "Introduction to the CoNLL-2005 shared task: Semantic role labeling". In: *Proceedings of the Ninth Conference on Computational Natural Language Learning*. Ed. by Ido Dagan and Dan Gildea. Ann Arbor, Michigan: Association for Computational Linguistics, June 29–May 30, 2005, pp. 152–164.

Carroll, John B. (1964). *Language and Thought*. New Jersey: Prentice-Hall.

Chan, Yee Seng and Ng, Hwee Tou (2008). "MAXSIM: A maximum similarity metric for machine translation evaluation". In: *Proceedings of 46th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*. Ed. by Kathleen McKeown, ohanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui. Columbus, Ohio: Association for Computational Linguistics, June 15–20, 2008, pp. 55–62.

Che, Wanxiang, Li, Zhenghua, and Liu, Ting (2010). "LTP: A Chinese language technology platform". In: *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations Volume*. Ed. by Yang Liu and Ting Liu. Beijing, China: Coling 2010 Organizing Committee, Aug. 23–27, 2010, pp. 13–16.

Chen, Tianqi and Guestrin, Carlos (2016). "XGBoost: A scalable tree boosting system". In: *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Ed. by Alex Smola, Charu Aggarwal, Dou Shen, and Rajeev Rastogi. KDD '16. San Francisco, California, USA: Association for Computing Machinery, Aug. 13–17, 2016, pp. 785–794.

Chiang, David (2007). "Hierarchical phrase-based translation". *Computational Linguistics*, 33(2) (June 2007), pp. 201–228.

Cho, Kyunghyun, Merrienboer, Bart van, Gulcehre, Caglar, Bahdanau, Dzmitry, Bougares, Fethi, Schwenk, Holger, and Bengio, Yoshua (2014). "Learning phrase representations using RNN encoder–decoder for statistical machine translation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Alessandro Moschitti, Bo Pang, and Walter Daelemans. Doha, Qatar: Association for Computational Linguistics, Oct. 25–29, 2014, pp. 1724–1734.

Chorowski, Jan K, Bahdanau, Dzmitry, Serdyuk, Dmitriy, Cho, Kyunghyun, and Bengio, Yoshua (2015). "Attention-based models for speech recognition". In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett. Curran Associates, Inc., pp. 577–585.

Church, K. and Gale, W. (1999). "Inverse document frequency (IDF): A measure of deviations from poisson". In: *Natural Language Processing Using Very Large Corpora*. Ed. by Susan Armstrong, Kenneth Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann, and David Yarowsky. Dordrecht: Springer Netherlands, pp. 283–295.

Church, Kenneth W. and Hovy, Eduard H. (1993). "Good applications for crummy machine translation". *Machine Translation*, 8(4) (Dec. 1993), pp. 239–258.

Church, Kenneth Ward and Hanks, Patrick (1990). "Word association norms, mutual information, and lexicography". *Computational linguistics*, 16(1), pp. 22–29.

Collins, Michael (2013). “Language modeling: Course notes for NLP”. course notes. Available at http://www.cs.columbia.edu/~mcollins/lm-spring2013.pdf. [Accessed on 20 May 2017]. New York, NY.

Collins, Michael and Duffy, Nigel (2001). “Convolution kernels for natural language”. In: *Advances in Neural Information Processing Systems*. Ed. by T. G. Dietterich, S. Becker, and Z. Ghahramani. Vancouver, British Columbia, Canada, Dec. 3–8, 2001, pp. 625–632.

Collobert, Ronan and Weston, Jason (2008). “A unified architecture for natural language processing: Deep neural networks with multitask learning”. In: *Proceedings of the 25th International Conference on Machine Learning*. Ed. by Andrew McCallum and Sam Roweis. ICML ’08. Helsinki, Finland: Association for Computing Machinery, July 5–9, 2008, pp. 160–167.

Collobert, Ronan, Weston, Jason, Bottou, Léon, Karlen, Michael, Kavukcuoglu, Koray, and Kuksa, Pavel (2011). “Natural language processing (almost) from scratch”. *Journal of Machine Learning Research*, 12(Aug), pp. 2493–2537.

Comelles, Elisabet, Atserias, Jordi, Arranz, Victoria, and Castellón, Irene (2012). “VERTa: Linguistic features in MT evaluation”. In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Istanbul, Turkey: European Language Resources Association (ELRA), May 23–25, 2012, pp. 3944–3950.

Comelles, Elisabet, Gimenez, Jesus, Marquez, Lluis, Castellon, Irene, and Arranz, Victoria (2010). “Document-level automatic MT evaluation based on discourse representations”. In: *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. Ed. by Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, and Omar Zaidan. Uppsala, Sweden: Association for Computational Linguistics, July 15–16, 2010, pp. 333–338.

Conrado, Merley, Pardo, Thiago, and Rezende, Solange (2013). “A Machine Learning Approach to Automatic Term Extraction using a Rich Feature Set”. In: *Proceedings of the 2013 NAACL HLT Student Research Workshop*. Ed. by Annie Louis, Richard Socher, Julia Hockenmaier, and Eric K. Ringger. Atlanta, Georgia: Association for Computational Linguistics, June 13, 2013, pp. 16–23.

Corston-Oliver, Simon, Gamon, Michael, and Brockett, Chris (2001). “A machine learning approach to the automatic evaluation of machine translation”. In: *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. Ed. by Norbert Reithinger and Giorgio Satta. ACL ’01. Toulouse, France: Association for Computational Linguistics, July 6–11, 2001, pp. 148–155.

Cortes, Corinna and Vapnik, Vladimir (1995). “Support-vector networks”. *Machine Learning*, 20(3) (Sept. 1995), pp. 273–297.

Corver, Norbert and Riemsdijk, Henk van (2001). *Semi-lexical categories: the function of content words and the content of function words*. Vol. 59. Berlin, New York: Walter de Gruyter.

Costa, Ângela, Ling, Wang, Luís, Tiago, Correia, Rui, and Coheur, Luísa (2015). “A linguistically motivated taxonomy for Machine Translation error analysis”. *Machine Translation*, 29(2) (June 2015), pp. 127–161.

Crossley, Scott A, Kyle, Kristopher, and McNamara, Danielle S (2016a). “The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality”. *Journal of Second Language Writing*, 32, pp. 1–16.

Crossley, Scott A, Kyle, Kristopher, and McNamara, Danielle S (2016b). “The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion”. *Behavior research methods*, 48(4), pp. 1227–1237.

Daille, Béatrice (2003). “Conceptual structuring through term variations”. In: *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18*. Ed. by Lori Levin, Takenobu Tokunaga, and Alessandro Lenci. MWE ’03. Sapporo, Japan: Association for Computational Linguistics, Jan. 12, 2003, pp. 9–16.

Daille, Béatrice and Morin, Emmanuel (2005). “French-English terminology extraction from comparable corpora”. In: *Proceedings of the Second International Joint Conference*. Ed. by Robert Dale, Kam-Fai Wong, Jian Su, and Oi Yee Kwong. Berlin, Heidelberg: Springer Berlin Heidelberg, Oct. 11–13, 2005, pp. 707–718.

De Marneffe, Marie-Catherine, MacCartney, Bill, Manning, Christopher D, et al. (2006). “Generating typed dependency parses from phrase structure parses”. In: *Proceedings of 5th Edition of the International Conference on Language Resources and Evaluation (LREC 2006)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk, and Daniel Tapias. Genoa,Italy: European Language Resources Association (ELRA), May 22–28, 2006, pp. 449–454.

Déjean, Hervé and Gaussier, Eric (2002). “Une nouvelle approche à l’extraction de lexiques bilingues à partir de corpus comparables”. *Lexicometrica, Alignement Lexical dans Les Corpus Multilingues*. Ed. by Jean Véronic. [Online] Available from http://lexicometrica.univ-paris3.fr/thema/thema6/Dejean.pdf. [Accessed on 23 December 2017], pp. 1–22.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). “Maximum likelihood from incomplete data via the EM algorithm”. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), pp. 1–38.

DeNero, John and Klein, Dan (2008). “The complexity of phrase alignment problems”. In: *Proceedings of 46th Annual Meeting of the Association for Computional Linguistics: Human Language Technologies*. Ed. by Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui. Columbus, Ohio: Association for Computational Linguistics, June 16–17, 2008, pp. 25–28.

Deng, Yonggang and Byrne, William (2005). “HMM word and phrase alignment for statistical machine translation”. In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Ed. by Chris Brew, Lee-Feng Chien, and Katrin Kirchhoff. Vancouver, British Columbia, Canada: Association for Computational Linguistics, Oct. 6–8, 2005, pp. 169–176.

Dickinson, Brian and Hu, Wei (2015). “Sentiment analysis of investor opinions on twitter”. *Social Networking*, 4(03), pp. 62–71.

Dinu, Georgiana and Baroni, Marco (2014). “Improving zero-shot learning by mitigating the hubness problem”. *CoRR*, abs/1412.6568.

Doddington, George (2002). “Automatic evaluation of machine translation quality using n-gram co-occurrence statistics”. In: *Proceedings of the Second International Conference on Human Language Technology Research*. Ed. by Mitchell Marcus and David Yarowsky. HLT ’02. San Diego, California: Defense Advanced Research Projects Agency, Mar. 24–27, 2002, pp. 138–145.

Domingos, Pedro (2012a). “A Few Useful Things to Know about Machine Learning”. *Commun. ACM*, 55(10) (Oct. 2012), pp. 78–87.

Domingos, Pedro (2012b). “A few useful things to know about machine learning”. *Commun. ACM*, 55(10) (Oct. 2012), pp. 78–87.

Douglas, Dan and Smith, Jan (1997). *Theoretical underpinnings of the test of spoken English revision project*. Available from https://www.ets.org/Media/Research/pdf/RM-97-02.pdf. [Accessed on 20 April 2018]. Princeton, New Jersey: Educational Testing Service.

Drugan, Joanna (2013). *Quality in professional translation: Assessment and improvement*. Vol. 9. London, New Delhi, New York, Sydney: Bloomsbury Academic.

Duchi, John, Hazan, Elad, and Singer, Yoram (2011). “Adaptive subgradient methods for online learning and stochastic optimization”. *Journal of Machine Learning Research*, 12(Jul), pp. 2121–2159.

Dugast, Daniel (1980). *La statistique lexicale*. Vol. 9. Genève: Éditions Slatkine.

Erdmann, Maike, Nakayama, Kotaro, Hara, Takahiro, and Nishio, Shojiro (2009). “Improving the extraction of bilingual terminology from wikipedia”. *CM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 5(4) (Oct. 2009), 31:1–31:17.

Esplà-Gomis, Miquel, Sánchez-Martínez, Felipe, and Forcada, Mikel (2015). “UAlacant word-level machine translation quality estimation system at WMT 2015”. In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, Pavel Pecina, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Hervé Saint-Amand, Carolina Scarton, Lucia Specia, and Marco Turchi. Lisbon, Portugal: Association for Computational Linguistics, Sept. 17–18, 2015, pp. 309–315.

Esplà-Gomis, Miquel, Sánchez-Martínez, Felipe, and Forcada, Mikel (2016). “UAlacant word-level and phrase-level machine translation quality estimation systems at WMT 2016”. In: *Proceedings of the First Conference on Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Rajan Chatterjee, Christian Federmann, Liane Guillou, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Aurelie Neveol, Mariana Neves, Pavel Pecina, Martin Popel, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Lucia Specia, Karin Verspoor, Joerg Tiedemann, and Marco Turchi. Berlin, Germany: Association for Computational Linguistics, Aug. 11–12, 2016, pp. 782–786.

Fahmi, Ismail, Bouma, Gosse, and Plas, Lonneke van der (2007). “Improving statistical method using known terms for automatic term extraction”. In: ed. by Peter-Arno Coppen, Hans van Halteren, and en Suzan Verberne. Nijmegen, Netherlands: University of Nijmegen, Dec. 7, 2007, pp. 1–8.

Farrús, Mireia, Costa-jussà, Marta R., Mariñ0, José, and Fonollosa, José A.R. (2010). “Linguistic-based evaluation criteria to identify statistical machine translation errors”. In: *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*. Ed. by François Yvon and Viggo Hansen. Saint-Raphaël, France: European Association for Machine Translation, May 27–28, 2010.

Faruqui, Manaal and Dyer, Chris (2014). “Improving Vector Space Word Representations Using Multilingual Correlation”. In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, Apr. 26–30, 2014, pp. 462–471.

Federico, Marcello, Bertoldi, Nicola, and Cettolo, Mauro (2008). “IRSTLM: an open source toolkit for handling large scale language models”. In: *INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Associ-*

ation, Brisbane, Australia, September 22-26, 2008. Brisbane, Australia, Sept. 22–26, 2008, pp. 1618–1621.

Federico, Marcello, Negri, Matteo, Bentivogli, Luisa, and Turchi, Marco (2014). "Assessing the impact of translation errors on machine translation quality with mixed-effects models". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bo Pang. Doha, Qatar: Association for Computational Linguistics, Oct. 25–29, 2014, pp. 1643–1653.

Felice, Mariano and Specia, Lucia (2012). "Linguistic features for quality estimation". In: *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Ed. by Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Montréal, Canada: Association for Computational Linguistics, June 7–8, 2012, pp. 96–103.

Ferrucci, David and Lally, Adam (2004). "UIMA: An architectural approach to unstructured information processing in the corporate research environment". *Natural Language Engineering*, 10(3-4), pp. 327–348.

Font-Llitjós, Ariadna, Carbonell, Jaime G, and Lavie, Alon (2005). "A framework for interactive and automatic refinement of transfer-based machine translation". In: ed. by Bente Maegaard, Viggo Hansen, Steven Krauwer, Gábor Prószéky, Harold Somers, and Gregor Thurmair. Budapest, Hungary: European Association for Machine Translation, May 30–31, 2005, pp. 87–96.

Formiga, Lluis, Marquez, Lluis, and Pujantell, Jaume (2013). "Real-life translation quality estimation for MT system selection". In: *Proceedings of Machine Translation Summit XIV*. Ed. by Andy Way, Khalil Sima'an, Mikel L. Forcada, Daniel Grasmick, and Heidi Depaetere. Nice,France: International Association for Machine Translation, Sept. 2–6, 2013, pp. 69–76.

Frantzi, Katerina T and Ananiadou, Sophia (1996). "Extracting nested collocations". In: *Proceedings of the 16th Conference on Computational linguistics-Volume 1*. Ed. by Jun'ich Tsujii. Association for Computational Linguistics. Copenhagen, Denmark, Aug. 5–9, 1996, pp. 41–46.

Fukunaga, Keinosuke and Hayes, Raymond (1989). "Effects of sample size in classifier design". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(8) (Aug. 1989), pp. 873–885.

Fung, Pascale and Yee, Lo Yuen (1998). "An IR approach for translating new words from nonparallel, comparable texts". In: *Proceedings of the 17th international conference on Computational linguistics-Volume 1*. Ed. by Christian Boitet and Pete Whitelock. Montreal, Quebec, Canada: Association for Computational Linguistics, Aug. 10–14, 1998, pp. 414–420.

Gaizauskas, Robert, Paramita, Monica Lestari, Barker, Emma, Pinnis, Marcis, Aker, Ahmet, and Solé, Marta Pahisa (2015). "Extracting bilingual terms from the Web". *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 21(2), pp. 205–236.

Gamon, Michael (2004). "Linguistic correlates of style: authorship classification with deep linguistic analysis features". In: *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*. Geneva, Switzerland: Association for Computational Linguistics, Aug. 23–27, 2004, pp. 611–617.

Gamon, Michael, Aue, Anthony, and Smets, Martine (2005). "Sentence-level MT evaluation without reference translations: beyond language modeling". In: *Proceedings of the 10th European Association for Machine Translation, EAMT 2005*. Ed. by Bente Maegaard, Viggo Hansen, Steven Krauwer, Gábor Prószéky,

Harold Somers, and Gregor Thurmair. Budapest, Hungry: European Association for Machine Translation, May 30–31, 2005, pp. 103–111.

Gandrabur, Simona and Foster, George (2003). “Confidence estimation for translation prediction”. In: *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003-Volume 4*. Ed. by Walter Daelemans and Miles Osborne. Edmonton, Canada: Association for Computational Linguistics, May 27–June 1, 2003, pp. 95–102.

Gaussier, Eric (2001). “General considerations on bilingual terminology extraction”. In: *Recent Advances in Computational Terminology*. Ed. by Didier Bourigault, Christian Jacquemin, and Marie-Claude L’Homme. Amsterdam: John Benjamins, pp. 167–183.

Giménez, Jesús and Màrquez, Lluís (2007). “Linguistic features for automatic evaluation of heterogenous MT systems”. In: *Proceedings of the Second Workshop on Statistical Machine Translation*. Ed. by Chris Callison-Burch, Philipp Koehn, Christof Monz, and Cameron Shaw Fordyce. StatMT ’07. Prague, Czech Republic: Association for Computational Linguistics, June 23, 2007, pp. 256–264.

Giménez, Jesús and Màrquez, Lluís (2009). “On the robustness of syntactic and semantic features for automatic MT evaluation”. In: *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Ed. by Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. StatMT ’09. Athens, Greece: Association for Computational Linguistics, Mar. 30–31, 2009, pp. 250–258.

Giménez, Jesús, Màrquez, Lluís, Comelles, Elisabet, Castellón, Irene, and Arranz, Victoria (2010). “Document-level automatic MT evaluation based on discourse representations”. In: *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. Ed. by Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, and Omar Zaidan. WMT ’10. Uppsala, Sweden: Association for Computational Linguistics, July 15–16, 2010, pp. 333–338.

Girardi, Christian, Bentivogli, Luisa, Farajian, Mohammad Amin, and Federico, Marcello (2014). “MT-EQuAl: a Toolkit for Human Assessment of Machine Translation Output”. In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*. Ed. by Jan Hajic and Junichi Tsujii. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, Aug. 23–29, 2014, pp. 120–123.

Goldberg, Yoav (2017). “Neural network methods for natural language processing”. *Synthesis Lectures on Human Language Technologies*, 10(1), pp. 1–309.

Goldberg, Yoav and Hirst, Graeme (2017). *Neural Network Methods in Natural Language Processing*. Morgan & Claypool.

Gomaa, Wael H and Fahmy, Aly A (2013). “A survey of text similarity approaches”. *International Journal of Computer Applications*, 68(13), pp. 13–18.

González-Rubio, Jesús, Sanchís, Alberto, and Casacuberta, Francisco (2012). “PRHLT submission to the wmt12 quality estimation task”. In: *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Ed. by Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Montréal, Canada: Association for Computational Linguistics, June 7–8, 2012, pp. 101–105.

Görög, A (2014a). “Quality evaluation today: the dynamic quality framework”. In: *Proceedings of the 36th Conference Translating and the Computer*. Ed. by Joao Esteves-Ferreira, Juliet Macan, Ruslan Mitkov, and Olaf-Michael Stefanov. London: AsLing: The International Association for Advancement in Language Technology, Nov. 27–28, 2014, pp. 155–164.

Görög, Attila (2014b). “Quantifying and benchmarking quality”. *Tradumàtica*, (12), pp. 443–454.

Gouws, Stephan, Bengio, Yoshua, and Corrado, Greg (2015). “BilBOWA: Fast bilingual distributed representations without word alignments”. In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. Ed. by Francis Bach and David Blei. ICML’15. Lille, France: JMLR.org, July 6–11, 2015, pp. 748–756.

Gouws, Stephan and Søgaard, Anders (2015). “Simple task-specific bilingual word embeddings”. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Joyce Chai and Anoop Sarkar. Denver, Colorado: Association for Computational Linguistics, May 31–June 5, 2015, pp. 1386–1390.

Graham, Yvette, Baldwin, Timothy, Moffat, Alistair, and Zobel, Justin (2017a). “Can machine translation systems be evaluated by the crowd alone”. *Natural Language Engineering*, 23(1), pp. 3–30.

Graham, Yvette, Ma, Qingsong, Baldwin, Timothy, Liu, Qun, Parra, Carla, and Scarton, Carolina (2017b). “Improving evaluation of document-level machine translation quality estimation”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Ed. by Phil Blunsom and Alexander Koller. Valencia, Spain: Association for Computational Linguistics, Apr. 3–7, 2017, pp. 356–361.

Graves, A., Mohamed, A. r., and Hinton, G. (2013). “Speech recognition with deep recurrent neural networks”. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. Ed. by Rabab Ward, Li Deng, Vikram Krishnamurthy, and Kostas Plataniotis. Vancouver, Canada: IEEE Xplore Digital Library, May 26–31, 2013, pp. 6645–6649.

Green, Paul E. and Srinivasan, V. (1978). “Conjoint Analysis in Consumer Research: Issues and Outlook”. *Journal of Consumer Research*, 5(2), pp. 103–123.

Green, Spence, Heer, Jeffrey, and Manning, Christopher D (2013). “The efficacy of human post-editing for language translation”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Ed. by Susanne Bødker, Steven Brewster, Patrick Baudisch, Michel Beaudouin-Lafon, and Wendy E. Mackay. Association for Computing Machinery. Paris, France, Apr. 22–May 2, 2013, pp. 439–448.

Guillou, Liane (2013). “Analysing lexical consistency in translation”. In: *Proceedings of the Workshop on Discourse in Machine Translation*. Ed. by Bonnie Webber, Andrei Popescu-Belis, Katja Markert, and Jörg Tiedemann. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 9, 2013, pp. 10–18.

Guiraud, Pierre (1954). *Les caractères statistiques du vocabulaire: essai de méthodologie*. Paris: Presses universitaires de France.

Gupta, Rohit, Orasan, Constantin, and Genabith, Josef van (2015). “ReVal: A simple and effective machine translation evaluation metric based on recurrent neural networks”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Ed. by Lluís Màrquez, Chris Callison-Burch, and Jian Su. Lisbon, Portugal: Association for Computational Linguistics, Sept. 17–21, 2015, pp. 1066–1072.

Guzmán, Francisco, Joty, Shafiq, Màrquez, Lluís, and Nakov, Preslav (2015). “Pairwise neural machine translation evaluation”. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Ed. by Chengqing Zong and Michael Strube. Beijing, China: Association for Computational Linguistics, July 25–31, 2015, pp. 805–814.

Guzmán, Francisco, Joty, Shafiq, Màrquez, Lluís, and Nakov, Preslav (2017). "Machine translation evaluation with neural networks". *Computer Speech & Language*, 45, pp. 180–200.

Hajič, Jan, Ciaramita, Massimiliano, Johansson, Richard, Kawahara, Daisuke, Martí, Maria Antònia, Màrquez, Lluís, Meyers, Adam, Nivre, Joakim, Padó, Sebastian, Štěpánek, Jan, Straňák, Pavel, Surdeanu, Mihai, Xue, Nianwen, and Zhang, Yi (2009). "The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages". In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL): Shared Task*. Ed. by Suzanne Stevenson and Xavier Carreras. CoNLL '09. Boulder, Colorado: Association for Computational Linguistics, June 4, 2009, pp. 1–18.

Hakami, H and Bollegala, Danushka (2017). "A classification approach for detecting cross-lingual biomedical term translations". *Natural Language Engineering*, 23(1), pp. 31–51.

Halliday, Michael Alexander Kirkwood and Hasan, Ruqaiya (2014). *Cohesion in english*. London and New York: Routledge.

Han, Aaron L-F, Wong, Derek F, Chao, Lidia S, He, Liangye, and Lu, Yi (2014). "Unsupervised quality estimation model for English to German translation and its application in extensive supervised evaluation". *The Scientific World Journal*, 2014. Article ID 760301.Available from http://dx.doi.org/10.1155/2014/760301. [Accessed on 10 May 2018].

Han, Aaron Li-Feng, Lu, Yi, Wong, Derek F., Chao, Lidia S., He, Liangye, and Xing, Junwen (2013). "Quality estimation for machine translation using the joint method of evaluation criteria and statistical modeling". In: *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Chris Callison-Burch, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Hervé Saint-Amand, Radu Soricut, and Lucia Specia. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 8–9, 2013, pp. 365–372.

Han, Aaron Li-Feng and Wong, Derek Fai (2016). "Machine translation evaluation: A survey". *CoRR*, abs/1605.04515. [Online] Available from http://arxiv.org/abs/1605.04515. [Accessed on 22 April 2018].

Hardmeier, Christian (2014). "Discourse in statistical machine translation". PhD thesis. Uppsala, Sweden: Department of Linguistics and Philology.

Hardmeier, Christian and Federico, Marcello (2010). "Modelling pronominal anaphora in statistical machine translation". In: *Proceedings of the 7th International Workshop on Spoken Language Translation*. Ed. by Joseph Mariani and Alex Waibel. Paris, France: Advanced Telecommunication Research Institute International (ATR), Dec. 2–3, 2010, pp. 283–289.

Hardmeier, Christian, Nivre, Joakim, and Tiedemann, Jörg (2012). "Tree kernels for machine translation quality estimation". In: *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Ed. by Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Montréal, Canada: Association for Computational Linguistics, June 7–8, 2012, pp. 109–113.

Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome (2009). *The Elements of Statistical Learning*. 2nd ed. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., p. 745.

Hatim, Basil and Mason, Ian (2014). *Discourse and the Translator*. London and New York: Routledge.

Heeringa, Wilbert Jan (2004). “Measuring dialect pronunciation differences using Levenshtein distance”. PhD thesis. Groningen, Netherlands: Centre for Language and Cognition, Jan. 2004.

Heid, Ulrich (1998). “A Linguistic Bootstrapping Approach to the Extraction of Term Candidates from German Text”. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 5(2), pp. 161–181.

Herdan, Gustav and Wijk, Nicolai v. (1960). *Type-token mathematics: A textbook of mathematical linguistics*. Vol. 4. The Hague: Mouton & Co.

Hewavitharana, Sanjika and Vogel, Stephan (2016). “Extracting parallel phrases from comparable data for machine translation”. *Natural Language Engineering*, 22(04), pp. 549–573.

Hinton, Geoffrey E., Srivastava, Nitish, Krizhevsky, Alex, Sutskever, Ilya, and Salakhutdinov, Ruslan (2012). “Improving neural networks by preventing co-adaptation of feature detectors”. *CoRR*, abs/1207.0580. [Online] Available from http://arxiv.org/abs/1207.0580. [Accessed on 25 April 2018].

Ho, Tin Kam (1995). “Random decision forests”. In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Montreal, Canada: IEEE Computer Society, Aug. 14–16, 1995, pp. 278–282.

Hokamp, Chris, Calixto, Iacer, Wagner, Joachim, and Zhang, Jian (2014). “Target-centric features for translation quality estimation”. In: *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Matouš Macháček, Christof Monz, Pavel Pecina, Matt Post, Hervé Saint-Amand, Radu Soricut, and Lucia Specia. Baltimore, Maryland, USA: Association for Computational Linguistics, June 26–27, 2014, pp. 329–334.

House, Juliane (2014). “Translation quality assessment: past and present”. In: *Translation: A Multidisciplinary Approach*. London: Palgrave Macmillan UK, pp. 241–264.

Hughes, Gordon (1968). “On the mean accuracy of statistical pattern recognizers”. *IEEE Transactions on Information Theory*, 14(1) (Jan. 1968), pp. 55–63.

Huot, Brian (1990). “Reliability, validity, and holistic scoring: What we know and what we need to know”. *College Composition and Communication*, 41(2), pp. 201–213.

Ilisei, Iustina, Inkpen, Diana, Corpas Pastor, Gloria, and Mitkov, Ruslan (2010). “Identification of Translationese: A Machine Learning Approach”. In: *Computational Linguistics and Intelligent Text Processing*. Ed. by Alexander Gelbukh. Berlin, Heidelberg: Springer Berlin Heidelberg, Mar. 21–27, 2010, pp. 503–511.

Imamura, Kenji (2002). “Application of translation knowledge acquired by hierarchical phrase alignment for pattern-based MT”. In: *Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation*. Ed. by Sergei Nirenburg, Teruko Mitamura, Eric Nyberg, Francis Bond, and Hiromi Nakaiwa. Keihanna, Japan: European Association for Machine Translation, Mar. 13–17, 2002, pp. 74–84.

Isenhour, Michelle and Kramlich, Gary (2008). “Holistic Grading: Are all Mistakes Created Equal?” *PRIMUS*, 18(5), pp. 441–448.

Isozaki, Hideki, Hirao, Tsutomu, Duh, Kevin, Sudoh, Katsuhito, and Tsukada, Hajime (2010). “Automatic evaluation of translation quality for distant language pairs”. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Ed. by Hang Li and Lluís Màrquez. Cambridge, MA: Association for Computational Linguistics, Oct. 9–11, 2010, pp. 944–952.

Jarvis, Scott (2002). “Short texts, best-fitting curves and new measures of lexical diversity”. *Language Testing*, 19(1), pp. 57–84.

Joty, Shafiq, Guzmán, Francisco, Màrquez, Lluís, and Nakov, Preslav (2017). “Discourse structure in machine translation evaluation”. *Computational Linguistics*, 43(4), pp. 683–722.

Jurafsky, Dan and Martin, James H (2008). “Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition”. In: 2nd ed. Prentice Hall series in artificial intelligence. New Jersey: Prentice Hall. Chap. N-grams.

Jurafsky, Dan and Martin, James H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. 2nd. Upper Saddle River, New Jersey: Pearson Prentice Hall.

Justeson, John S and Katz, Slava M (1995). “Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text”. *Natural language engineering*, 1(01), pp. 9–27.

Kageura, Kyo and Umino, Bin (1996). “Methods of automatic term recognition: A review”. *Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication*, 3(2), pp. 259–289.

Kaljahi, Rasoul, Foster, Jennifer, and Roturier, Johann (2014a). “Syntax and semantics in quality estimation of machine translation”. In: *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Ed. by Dekai Wu, Marine Carpuat, Xavier Carreras, and Eva Maria Vecchi. Doha, Qatar: Association for Computational Linguistics, Oct. 25, 2014, pp. 67–77.

Kaljahi, Rasoul, Foster, Jennifer, Roturier, Johann, and Rubino, Raphael (2014b). “Quality estimation of English-French machine translation: A detailed study of the role of syntax”. In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Ed. by Junichi Tsujii and Jan Hajic. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, Aug. 23–29, 2014, pp. 2052–2063.

Kamp, Hans and Reyle, Uwe (2013). *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory*. Vol. 42. Dordrecht: Springer.

Kanal, Laveen and Chandrasekaran, B. (1971). “On dimensionality and sample size in statistical pattern classification”. *Pattern Recognition*, 3(3), pp. 225–234.

Kang, Byeong-Kwu, Chang, Bao-Bao, Chen, Yi-Rong, and Yu, Shi-Wen (2005). “Extracting terminologically relevant collocations in the translation of Chinese monograph”. In: *Proceedings of the International Conference on Natural Language Processing*. Ed. by Robert Dale, Kam-Fai Wong, Jian Su, and Oi Yee Kwong. Jeju Island, Korea: Springer, Oct. 11–13, 2005, pp. 1017–1028.

Karoubi, Behrouz (2016). “Translation quality assessment demystified”. *Babel*, 62(2), pp. 253–277.

Katz, Slava (1987). “Estimation of probabilities from sparse data for the language model component of a speech recognizer”. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(3) (Mar. 1987), pp. 400–401.

Kendall, Maurice G (1938). “A new measure of rank correlation”. *Biometrika*, 30(1/2), pp. 81–93.

Kim, Hyun and Lee, Jong-Hyeok (2016a). “A recurrent neural networks approach for estimating the quality of machine translation output”. In: *Proceedings of the 2016*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Ed. by Kevin Knight, Ani Nenkova, and Owen Rambow. San Diego, California: Association for Computational Linguistics, June 12–17, 2016, pp. 494–498.

Kim, Hyun and Lee, Jong-Hyeok (2016b). "Recurrent neural network based translation quality estimation". In: *Proceedings of the First Conference on Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Rajan Chatterjee, Christian Federmann, Liane Guillou, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Aurelie Neveol, Mariana Neves, Pavel Pecina, Martin Popel, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Lucia Specia, Karin Verspoor, Joerg Tiedemann, and Marco Turchi. Berlin, Germany: Association for Computational Linguistics, Aug. 11–12, 2016, pp. 787–792.

Kim, Hyun, Lee, Jong-Hyeok, and Na, Seung-Hoon (2017). "Predictor-Estimator using Multilevel Task Learning with Stack Propagation for Neural Quality Estimation". In: *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 8, 2017, pp. 562–568.

Kim, J-D, Ohta, Tomoko, Tateisi, Yuka, and Tsujii, Jun'ichi (2003). "GENIA Corpus–A semantically annotated corpus for bio-textmining". *Bioinformatics*, 19(suppl 1) (July 2003), pp. i180–i182.

Kim, Taehong, Hwang, Myunggwon, Hwang, Mi-Nyeong, Song, Sa-kwang, Jeong, Do-Heon, and Jung, Hanmin (2015). "Translation of technical terminologies between English and Korean based on textual big data". *Software: Practice and Experience*, 45(8), pp. 1115–1126.

King, Jonathan and Just, Marcel Adam (1991). "Individual differences in syntactic processing: The role of working memory". *Journal of Memory and Language*, 30(5), pp. 580–602.

Kingma, Diederik P. and Ba, Jimmy (2014). "Adam: A Method for Stochastic Optimization". *CoRR*, abs/1412.6980. [Online] Available from http://arxiv.org/abs/1412.6980. [Accessed on 23 April 2018].

Kirchhoff, Katrin, Capurro, Daniel, and Turner, Anne (2012). "Evaluating user preferences in machine translation using conjoint analysis". In: *Proceedings of the 16th International Conference of the European Association for Machine Translation (EAMT)*. Ed. by Mauro Cettolo, Marcello Federico, Lucia Specia, and Andy Way. Trento, Italy: European Association for Machine Translation, May 28–30, 2012, pp. 119–126.

Klebanov, Beata Beigman and Flor, Michael (2013). "Associative texture is lost in translation". In: *Proceedings of the Workshop on Discourse in Machine Translation*. Ed. by Bonnie Webber, Katja Markert, Andrei Popescu-Belis, and Jörg Tiedemann. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 9, 2013, pp. 27–32.

Klein, Dan and Manning, Christopher D. (2003). "Accurate unlexicalized parsing". In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*. Ed. by Erhard W. Hinrichs and Dan Roth. ACL '03. Sapporo, Japan: Association for Computational Linguistics, July 7–12, 2003, pp. 423–430.

Knoth, Petr, Schmidt, Marek, Smrz, Pavel, and Zdráhal, Zdenek (2009). "Towards a framework for comparing automatic term recognition methods". In: ed. by Pavel Smrž, Sylva Otáhalová, Marek Schmidt, and Jana Slámová. Bratislava, SK: Faculty of Informatics and Information Technology Slovak University of Technology in Bratislava, Feb. 4–6, 2009, pp. 83–94.

Koehn, Philipp (2009). *Statistical machine translation*. Cambridge and New York: Cambridge University Press.

Koehn, Philipp and Monz, Christof (2006). "Manual and automatic evaluation of machine translation between European languages". In: *Proceedings of the Workshop on Statistical Machine Translation*. Ed. by Philipp Koehn and Christof Monz. StatMT '06. New York City, New York: Association for Computational Linguistics, June 8–9, 2006, pp. 102–121.

Koehn, Philipp, Och, Franz Josef, and Marcu, Daniel (2003). "Statistical phrase-based translation". In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*. Ed. by Marti Hearst and Mari Ostendorf. NAACL '03. Edmonton, Canada: Association for Computational Linguistics, May 27–June 1, 2003, pp. 48–54.

Kohavi, Ron (1995). "A Study of cross-validation and bootstrap for accuracy estimation and model selection". In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*. Ed. by C. Raymond Perrault and Chris S. Mellish. IJCAI'95. Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc., Aug. 20–25, 1995, pp. 1137–1143.

Koponen, Maarit, Aziz, Wilker, Ramos, Luciana, and Specia, Lucia (2012). "Post-editing time as a measure of cognitive effort". *Proceedings of AMTA 2012 Workshop on Post-editing Technology and Practice (WPTP)* (Oct. 28, 2012), pp. 11–20.

Kosub, Sven (2016). "A note on the triangle inequality for the Jaccard distance". *CoRR*, abs/1612.02696. [Online] Available from http://arxiv.org/abs/1612.02696. [Accessed on 24 April 2018].

Krehbiel, Timothy C. (2004). "Correlation Coefficient Rule of Thumb". *Decision Sciences Journal of Innovative Education*, 2(1) (Jan. 16, 2004), pp. 97–100.

Kreutzer, Julia, Schamoni, Shigehiko, and Riezler, Stefan (2015). "Quality estimation from ScraTCH (QUETCH): Deep learning for word-level translation Quality Estimation". In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, Pavel Pecina, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Hervé Saint-Amand, Carolina Scarton, Lucia Specia, and Marco Turchi. Lisbon, Portugal: Association for Computational Linguistics, Sept. 17–18, 2015, pp. 316–322.

Kruskal, Joseph B. (1983). "An overview of sequence comparison: Time warps, string edits, and macromolecules". *SIAM Review*, 25(2), pp. 201–237.

Kulesza, Alex and Shieber, Stuart M (2004). "A learning approach to improving sentence-level MT evaluation". In: *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*. Baltimore, Maryland, USA, Oct. 4–6, 2004, pp. 75–84.

Lafferty, John, McCallum, Andrew, and Pereira, Fernando CN (2001). "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". In: *Proceedings of the Eighteeth International Conference on Machine Learning*. Ed. by Carla Brodley and Andrea Danyluk. Massachusetts, USA, June 28–July 1, 2001.

Lagarda, A.-L., Alabau, V., Casacuberta, F., Silva, R., and Díaz-de-Liaño, E. (2009). "Statistical post-editing of a rule-based machine translation system". In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics,*

Companion Volume: Short Papers. Ed. by Mari Ostendorf, Michael Collins, Shri Narayanan, Douglas W. Oard, and Lucy Vanderwende. NAACL-Short '09. Boulder, Colorado: Association for Computational Linguistics, May 31–June 5, 2009, pp. 217–220.

Lample, Guillaume, Ballesteros, Miguel, Subramanian, Sandeep, Kawakami, Kazuya, and Dyer, Chris (2016). "Neural Architectures for Named Entity Recognition". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Kevin Knight, Ani Nenkova, and Owen Rambow. San Diego, California: Association for Computational Linguistics, June 12–17, 2016, pp. 260–270.

Landauer, Thomas K, Foltz, Peter W, and Laham, Darrell (1998). "An introduction to latent semantic analysis". *Discourse processes*, 25(2-3), pp. 259–284.

Langlois, David (2015). "LORIA system for the WMT15 quality estimation shared task". In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, Pavel Pecina, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Hervé Saint-Amand, Carolina Scarton, Lucia Specia, and Marco Turchi. Lisbon, Portugal: Association for Computational Linguistics, Sept. 17–18, 2015, pp. 323–329.

Langlois, David, Raybaud, Sylvain, and Smaïli, Kamel (2012). "LORIA system for the wmt12 quality estimation shared task". In: *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Ed. by Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Montréal, Canada: Association for Computational Linguistics, June 7–8, 2012, pp. 111–116.

Lapshinova-Koltunski, Ekaterina (2015). "Exploration of inter- and intralingual variation of discourse phenomena". In: *Proceedings of the Second Workshop on Discourse in Machine Translation*. Ed. by Bonnie Webber, Marine Carpuat, Andrei Popescu-Belis, and Christian Hardmeier. Lisbon, Portugal: Association for Computational Linguistics, Sept. 17, 2015, pp. 158–167.

Lê, Sébastien, Josse, Julie, Husson, François, et al. (2008). "FactoMineR: An R package for multivariate analysis". *Journal of Statistical Software*, 25(1), pp. 1–18.

Lemaître, Guillaume, Nogueira, Fernando, and Aridas, Christos K. (2017). "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning". *Journal of Machine Learning Research*, 18(17), pp. 1–5.

Leusch, Gregor, Ueffing, Nicola, and Ney, Hermann (2006). "CDER: Efficient MT evaluation using block movements". In: *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*. Ed. by Frank Keller and Gabor Proszeky. Trento, Italy: Association for Computational Linguistics, Apr. 5–6, 2006, pp. 241–248.

Levenshtein, Vladimir I (1966). "Binary codes capable of correcting deletions, insertions, and reversals". 10(8) (Feb. 1966), pp. 707–710.

Levy, Roger and Manning, Christopher D. (2003). "Is it harder to parse Chinese, or the Chinese treebank?" In: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Sapporo, Japan: Association for Computational Linguistics, July 7–12, 2003, pp. 439–446.

Li, Defeng (2006). "Making translation testing more teaching-oriented: A case study of translation testing in China". *Meta: Journal des traducteurs/Meta: Translators' Journal*, 51(1), pp. 72–88.

Li, Lishuang, Dang, Yanzhong, Zhang, Jing, and Li, Dan (2012a). “Domain term extraction based on conditional random fields combined with active learning strategy”. *Journal of Information & Computational Science*, 9(7), pp. 1931–1940.

Li, Tianyu, Chubak, Pirooz, Lakshmanan, Laks VS, and Pottinger, Rachel (2012b). “Efficient extraction of ontologies from domain specific text corpora”. In: *CIKM ’12: Proceedings of the 21st ACM international conference on Information and Knowledge Management*. Ed. by Xuewen Chen, Guy Lebanon, Haixun Wang, and Mohammed J.Zaki. Maui, Hawaii, USA: Association for Computing Machinery, Oct. 29–Nov. 2, 2012, pp. 1537–1541.

Li, Yang and Yang, Tao (2018). “Word embedding for understanding natural language: A survey”. In: *Guide to Big Data Applications*. Ed. by S. Srinivasan. Cham, Switzerland: Springer International Publishing, pp. 83–104.

Li, Yujian and Liu, Bo (2007). “A normalized Levenshtein distance metric”. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 29(6), pp. 1091–1095.

Lin, Chin-Yew (2004). “ROUGE: A package for automatic evaluation of summaries”. In: *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*. Ed. by Marie-Francine Moens and Stan Szpakowicz. Barcelona, Spain: Association for Computational Linguistics, July 25–26, 2004, pp. 74–81.

Lin, Chin-Yew and Och, Franz Josef (2004). “Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics”. In: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Ed. by Donia Scott, Walter Daelemans, and Marilyn Walker. ACL ’04. Barcelona, Spain: Association for Computational Linguistics, July 21–26, 2004, pp. 605–612.

Liu, Bang, Zhang, Ting, Niu, Di, Lin, Jinghong, Lai, Kunfeng, and Xu, Yu (2018). “Matching Long Text Documents via Graph Convolutional Networks”. *CoRR*, abs/1802.07459.

Liu, Ding and Gildea, Daniel (2005). “Syntactic features for evaluation of machine translation”. In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ed. by Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss. Ann Arbor, Michigan: Association for Computational Linguistics, June 29, 2005, pp. 25–32.

Liu, Yang, Liu, Qun, and Lin, Shouxun (2005). “Log-linear models for word alignment”. In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Ed. by Kevin Knight, Hwee Tou Ng, and Kemal Oflazer. ACL ’05. Ann Arbor, Michigan: Association for Computational Linguistics, June 25–30, 2005, pp. 459–466.

Liu, Yang, Xia, Tian, Xiao, Xinyan, and Liu, Qun (2009). “Weighted alignment matrices for statistical machine translation”. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Ed. by Philipp Koehn and Rada Mihalcea. Singapore: Association for Computational Linguistics, Aug. 6–7, 2009, pp. 1017–1026.

Lommel, Arle (2016). “Blues for BLEU : Reconsidering the validity of reference-based MT evaluation”. In: *Proceedings of the LREC 2016 Workshop “Translation Evaluation – From Fragmented Tools and Data Sets to an Integrated Ecosystem"*. Ed. by Georg Rehm, Aljoscha Burchardt, Ondřej Bojar, Christian Dugast, Marcello Federico, Josef van Genabith, Barry Haddow, Jan Hajič, Kim Harris, Philipp Koehn, Matteo Negri, Martin Popel, Lucia Specia, Marco Turchi, and Hans Uszkoreit. Portorož, Slovenia: European Language Resources Association (ELRA), May 24, 2016, pp. 63–70.

Lommel, Arle (2018). “Metrics for translation quality assessment: A case for standardising error typologies”. In: *Translation Quality Assessment: From Principles to Practice*. Ed. by Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty. Cham, Switzerland: Springer International Publishing, pp. 111–127.

Lommel, Arle, Uszkoreit, Hans, and Burchardt, Aljoscha (2014a). “Multidimensional quality metrics (MQM) : A framework for declaring and describing translation quality metrics”. *Tradumàtica*, (12), pp. 455–463.

Lommel, Arle, Uszkoreit, Hans, and Burchardt, Aljoscha (2014b). “Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics”. *Tradumàtica*, (12), pp. 455–463.

Lommel, Arle Richard, Burchardt, Alojscha, and Uszkoreit, Hans (2013). “Multidimensional Quality Metrics: A Flexible System for Assessing Translation Quality”. In: ed. by Ivelina Nikolova. Vol. 35. London: AsLing: The International Association for Advancement in Language Technology, Nov. 28–29, 2013.

Lu, Bin and Tsou, Benjamin K. (2009). “Towards Bilingual Term Extraction in Comparable Patents”. In: *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*. Ed. by Olivia Kwong. Hong Kong: City University of Hong Kong, Dec. 3–5, 2009, pp. 755–762.

Luong, Ngoc Quang, Besacier, Laurent, and Lecouteux, Benjamin (2014). “LIG system for word level QE task at WMT14”. In: *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Matouš Macháček, Christof Monz, Pavel Pecina, Matt Post, Hervé Saint-Amand, Radu Soricut, and Lucia Specia. Baltimore, Maryland, USA: Association for Computational Linguistics, June 26–27, 2014, pp. 335–341.

Luong, Ngoc Quang, Lecouteux, Benjamin, and Besacier, Laurent (2013). “LIG system for WMT13 QE Task: Investigating the usefulness of features in word confidence estimation for MT”. In: *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Chris Callison-Burch, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Hervé Saint-Amand, Radu Soricut, and Lucia Specia. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 8–9, 2013, pp. 386–391.

Luong, Ngoc-Quang, Besacier, Laurent, and Lecouteux, Benjamin (2015a). “Towards accurate predictors of word quality for machine translation: Lessons learned on French–English and English–Spanish systems”. *Data and Knowledge Engineering*, 96-97(2) (Mar.–May 2015). Knowledge and Systems Engineering- KSE 2013, pp. 32–42.

Luong, Thang, Pham, Hieu, and Manning, Christopher D. (2015b). “Effective approaches to attention-based neural machine translation”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Ed. by Lluís Màrquez, Chris Callison-Burch, and Jian Su. Lisbon, Portugal: Association for Computational Linguistics, Sept. 17–21, 2015, pp. 1412–1421.

Ma, Xuezhe and Hovy, Eduard (2016–2018). “End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Antal van den Bosch, Katrin Erk, and Noah A. Smith. Berlin, Germany: Association for Computational Linguistics, Aug. 7, 2016–Aug. 12, 2018, pp. 1064–1074.

Maaten, Laurens van der and Hinton, Geoffrey (2008). “Visualizing data using t-SNE”. *Journal of Machine Learning Research*, 9(Nov), pp. 2579–2605.

MacCartney, Bill, Galley, Michel, and Manning, Christopher D. (2008). “A phrase-based alignment model for natural language inference”. In: *Proceedings of*

the 2008 Conference on Empirical Methods in Natural Language Processing. Ed. by Mirella Lapata and Hwee Tou Ng. Honolulu, Hawaii: Association for Computational Linguistics, Oct. 25–27, 2008, pp. 802–811.

Mann, William C and Thompson, Sandra A (1988). "Rhetorical structure theory: Toward a functional theory of text organization". *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3), pp. 243–281.

Manning, Christopher D and Schütze, Hinrich (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press, May 1999.

Manning, Christopher D., Surdeanu, Mihai, Bauer, John, Finkel, Jenny, Bethard, Steven J., and McClosky, David (2014). "The Stanford CoreNLP natural language processing toolkit". In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Ed. by Kalina Bontcheva and Zhu Jingbo. Baltimore, Maryland, USA: Association for Computational Linguistics, June 23–24, 2014, pp. 55–60.

Martins, André F. T., Astudillo, Ramón, Hokamp, Chris, and Kepler, Fabio (2016). "Unbabel's participation in the WMT16 word-level translation quality estimation shared task". In: *Proceedings of the First Conference on Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Rajan Chatterjee, Christian Federmann, Liane Guillou, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Aurelie Neveol, Mariana Neves, Pavel Pecina, Martin Popel, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Lucia Specia, Karin Verspoor, Joerg Tiedemann, and Marco Turchi. Berlin, Germany: Association for Computational Linguistics, Aug. 11–12, 2016, pp. 806–811.

Martins, André F. T., Kepler, Fabio, and Monteiro, Jose (2017). "Unbabel's Participation in the WMT17 Translation Quality Estimation Shared Task". In: *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 8, 2017, pp. 569–574.

Marzal, A. and Vidal, E. (1993). "Computation of normalized edit distance and applications". *IEEE transactions on pattern analysis and machine intelligence*, 15(9) (Sept. 1993), pp. 926–932.

Matsuo, Yutaka and Ishizuka, Mitsuru (2004). "Keyword extraction from a single document using word co-occurrence statistical information". *International Journal on Artificial Intelligence Tools*, 13(01), pp. 157–169.

Melamed, I. Dan (1995). "Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons". In: *Proceedings of the Third Workshop on Very Large Corpora*. Ed. by David Yarowsky and Kenneth Church. Cambridge, Massachusetts, USA: Association for Computational Linguistics, June 30, 1995, pp. 184–198.

Melby, Alan K. (2015). "QT21: A new era for translators and the computer". In: *Proceedings of the 37th Conference Translating and the Computer*. Ed. by Joao Esteves-Ferreira, Juliet Macan, Ruslan Mitkov, and Olaf-Michael Stefanov. Westminster, London: AsLing: The International Association for Advancement in Language Technology, Nov. 26–27, 2015, pp. 1–11.

Meyer, Thomas, Popescu-Belis, Andrei, Hajlaoui, Najeh, and Gesmundo, Andrea (2012). "Machine translation of labeled discourse connectives". In: *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA)*. Ed. by Alon Lavie, George Foster, Mike Dillinger, Ray Flournoy, Nick Bemish, and Chuck Simmons. San Diego, California, USA: Association for Machine Translation in the Americas, Oct. 28–Nov. 1, 2012, pp. 11–20.

Meyer, Thomas and Webber, Bonnie (2013). “Implicitation of discourse connectives in (machine) translation”. In: *Proceedings of the Workshop on Discourse in Machine Translation*. Ed. by Bonnie Webber, Andrei Popescu-Belis, Katja Markert, and Jörg Tiedemann. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 9, 2013, pp. 19–26.

Mikolov, Tomas, Chen, Kai, Corrado, Greg, and Dean, Jeffrey (2013a). “Efficient Estimation of Word Representations in Vector Space”. *CoRR*, abs/1301.3781. [Online] Available from http://arxiv.org/abs/1301.3781. [Accessed on 23 April 2018].

Mikolov, Tomas, Le, Quoc V, and Sutskever, Ilya (2013b). “Exploiting similarities among languages for machine translation”. *arXiv preprint arXiv:1309.4168*.

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg S, and Dean, Jeff (2013c). “Distributed representations of words and phrases and their compositionality”. In: *Advances in Neural Information Processing Systems 26*. Ed. by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger. Montréal Canada: Curran Associates, Inc., Dec. 5–10, 2013, pp. 3111–3119.

Miller, Alan (2002). *Subset selection in regression*. London & New York: Chapman and Hall Press/CRC.

Miller, George A, Beckwith, Richard, Fellbaum, Christiane, Gross, Derek, and Miller, Katherine J (1990). “Introduction to WordNet: An on-line lexical database”. *International Journal of Lexicography*, 3(4), pp. 235–244.

Mohanty, Pradeep Kumar and Kumar, Patel Saroj (2015). *Basic statistics*. 1st ed. Springer Texts in Statistics. Jodhpur and Delhi, India: Scientific Publishers.

Mohri, Mehryar (1996). “On some applications of finite-state automata theory to natural langauge processing”. *Natural Language Engineering*, 2(1), pp. 61–80.

Moore, Robert C. (2005). “Association-based bilingual word alignment”. In: *Proceedings of Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*. Ed. by Philipp Koehn, Joel Martin, Rada Mihalcea, Christof Monz, and Ted Pederse. Ann Arbor, Michigan: Association for Computational Linguistics, June 29–30, 2005, pp. 1–8.

Moore, Robert C., Yih, Wen-tau, and Bode, Andreas (2006). “Improved discriminative bilingual word alignment”. In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. Ed. by Nicoletta Calzolari, Claire Cardie, and Pierre Isabelle. ACL-44. Sydney, Australia: Association for Computational Linguistics, July 17–21, 2006, pp. 513–520.

Moreau, Erwan and Rubino, Raphael (2013). “An approach using style classification features for quality estimation”. In: *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Chris Callison-Burch, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Hervé Saint-Amand, Radu Soricut, and Lucia Specia. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 8–9, 2013, pp. 429–434.

Moreau, Erwan and Vogel, Carl (2012). “Quality estimation: an experimental study using unsupervised similarity measures”. In: *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Ed. by Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Montréal, Canada: Association for Computational Linguistics, June 7–8, 2012, pp. 120–126.

Moreau, Erwan and Vogel, Carl (2014). “Limitations of MT quality estimation supervised systems: The tails prediction problem”. In: *Proceedings of COLING*

2014, the 25th International Conference on Computational Linguistics: Technical Papers. Ed. by Junichi Tsujii and Jan Hajic. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, Aug. 23–29, 2014, pp. 2205–2216.

Morris, Andrew Cameron, Maier, Viktoria, and Green, Phil (2004). "From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition". In: *Proceedings of the Eighth International Conference on Spoken Language Processing, INTERSPEECH 2004*. Ed. by Soon Hyob Kim, Sang-Oak Lee, and Yung-Hwan Oh. Jeju Island, Korea: International Speech Communication Association, Oct. 4–8, 2004, pp. 2765–2768.

Morris, Jane and Hirst, Graeme (1991). "Lexical cohesion computed by thesaural relations as an indicator of the structure of text". *Computational Linguistics*, 17(1) (Mar. 1991), pp. 21–48.

Moschitti, Alessandro (2006). "Efficient Convolution Kernels for Dependency and Constituent Syntactic Trees". In: *Machine Learning: ECML 2006*. Ed. by Johannes Fürnkranz, Tobias Scheffer, and Myra Spiliopoulou. Berlin, Heidelberg: Springer Berlin Heidelberg, Sept. 22–26, 2006, pp. 318–329.

Mucciardi, A. and Gose, E. (1971). "A Comparison of seven techniques for choosing subsets of pattern recognition properties". *IEEE Transactions on Computers*, C-20(9) (Sept. 1971), pp. 1023–1031.

Munday, Jeremy (2016a). *Introducing Translation Studies: Theories and Applications*. 4th ed. Routledge.

Munday, Jeremy (2016b). *Introducing translation studies: Theories and applications*. 4th ed. Milton Park & New York: Routledge.

Neubig, Graham, Watanabe, Taro, Sumita, Eiichiro, Mori, Shinsuke, and Kawahara, Tatsuya (2011). "An unsupervised model for joint phrase alignment and extraction". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Dekang Lin, Yuji Matsumoto, and Rada Mihalcea. Portland, Oregon, USA: Association for Computational Linguistics, June 19–24, 2011, pp. 632–641.

Nießen, Sonja, Och, Franz Josef, Leusch, Gregor, Ney, Hermann, et al. (2000). "An evaluation tool for machine translation: Fast evaluation for MT research." In: *Proceedings of LREC 2000 2nd International Conference on Language Resources & Evalaution*. Ed. by M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, and G. Stainhauer. Athens, Greece: European Language Resources Association (ELRA), May 31–June 2, 2000.

Novák, Michal, Oele, Dieke, and Noord, Gertjan van (2015). "Comparison of coreference resolvers for deep syntax translation". In: *Proceedings of the Second Workshop on Discourse in Machine Translation*. Ed. by Bonnie Webber, Marine Carpuat, Andrei Popescu-Belis, and Christian Hardmeier. Lisbon, Portugal: Association for Computational Linguistics, Sept. 17, 2015, pp. 17–23.

Och, Franz Josef and Ney, Hermann (2003). "A systematic comparison of various statistical alignment models". *Computational Linguistics*, 29(1), pp. 19–51.

Och, Franz Josef, Tillmann, Christoph, Ney, Hermann, et al. (1999). "Improved alignment models for statistical machine translation". In: *Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. Ed. by Pascale Fung and Joe Zhou. College Park, MD,USA: Association for Computational Linguistics, June 21–22, 1999, pp. 20–28.

Okita, Tsuyoshi, Rubino, Raphaël, and Genabith, Josef van (2012). "Sentence-level quality estimation for MT system combination". In: *Proceedings of the*

Second Workshop on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid MT. Ed. by Josef van Genabith, Toni Badia, Christian Federmann, Maite Melero, Marta R. Costa-jussà, and Tsuyoshi Okita. Mumbai, India: The COLING 2012 Organizing Committee, Dec. 15, 2012, pp. 55–64.

Oliver, Antoni (2017). "A system for terminology extraction and translation equivalent detection in real time". *Machine Translation*, 31(3) (Sept. 2017), pp. 147–161.

Paetzold, Gustavo and Specia, Lucia (2016). "SimpleNets: Quality estimation with resource-light neural networks". In: *Proceedings of the First Conference on Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Rajan Chatterjee, Christian Federmann, Liane Guillou, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Aurelie Neveol, Mariana Neves, Pavel Pecina, Martin Popel, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Lucia Specia, Karin Verspoor, Joerg Tiedemann, and Marco Turchi. Berlin, Germany: Association for Computational Linguistics, Aug. 11–12, 2016, pp. 812–818.

Palmer, Martha, Gildea, Daniel, and Kingsbury, Paul (2005). "The proposition bank: An annotated corpus of semantic roles". *Computational linguistics*, 31(1), pp. 71–106.

Papineni, Kishore, Roukos, Salim, Ward, Todd, and Zhu, Wei-Jing (2002). "BLEU: A method for automatic evaluation of machine translation". In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Ed. by Eugene Charniak and Dekang Lin. ACL '02. Philadelphia, Pennsylvania: Association for Computational Linguistics, July 7–12, 2002, pp. 311–318.

Park, Youngja, Byrd, Roy J, and Boguraev, Branimir K (2002). "Automatic glossary extraction: Beyond terminology identification". In: *Proceedings of the 19th International Conference on Computational Linguistics-Volume 1*. Ed. by Shu-Chuan Tseng. COLING '02. Taipei, Taiwan: Association for Computational Linguistics, Aug. 24–Sept. 1, 2002, pp. 1–7.

Payne, Thomas Edward (1997). *Describing morphosyntax: a guide for field linguists*. Cambridge: Cambridge University Press, Oct. 1997.

Pazienza, Maria Teresa, Pennacchiotti, Marco, and Zanzotto, Fabio Massimo (2005). "Terminology extraction: An analysis of linguistic and statistical approaches". In: *Knowledge Mining: Proceedings of the NEMIS 2004 Final Conference*. Ed. by Spiros Sirmakessis. Berlin, Heidelberg: Springer Berlin Heidelberg. Chap. 20, pp. 255–279.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). "Scikit-learn: Machine learning in python". *Journal of Machine Learning Research*, 12(Oct), pp. 2825–2830.

Pennington, Jeffrey, Socher, Richard, and Manning, Christopher (2014). "Glove: Global Vectors for word representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Alessandro Moschitti, Bo Pang, and Walter Daelemans. Doha, Qatar: Association for Computational Linguistics, Oct. 25–29, 2014, pp. 1532–1543.

Pérez-Paredes, Pascual and Sánchez-Tornel, María (2014). "Adverb use and language proficiency in young learners' writing". *International Journal of Corpus Linguistics*, 19(2), pp. 178–200.

Peters, Matthew, Neumann, Mark, Iyyer, Mohit, Gardner, Matt, Clark, Christopher, Lee, Kenton, and Zettlemoyer, Luke (2018). "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Ed. by Heng Ji and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, June 1–6, 2018, pp. 2227–2237.

Petrov, Slav, Das, Dipanjan, and McDonald, Ryan (2012–2013). "A universal part-of-speech tagset". In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Istanbul, Turkey: European Language Resources Association (ELRA), May 23, 2012–May 25, 2013, pp. 2089–2096.

Pfanzagl, Johann (1994). *Parametric statistical theory*. Berlin and New York: Walter de Gruyter, pp. 207–208.

Pighin, Daniele, González, Meritxell, and Màrquez, Lluís (2012). "The UPC submission to the wmt 2012 shared task on quality estimation". In: *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Ed. by Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Montréal, Canada: Association for Computational Linguistics, June 7–8, 2012, pp. 124–129.

Pitler, Emily and Nenkova, Ani (2009). "Using syntax to disambiguate explicit discourse connectives in text". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Ed. by Keh-Yih Su, Jian Su, Janyce Wiebe, and Haizhou Li. Suntec, Singapore: Association for Computational Linguistics, Aug. 2–7, 2009, pp. 13–16.

Pitman, Jim and Yor, Marc (1997). "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator". *The Annals of Probability*, 25(2), pp. 855–900.

Popovic, Maja (2012). "Morpheme- and POS-based IBM1 and language model scores for translation quality estimation". In: *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Ed. by Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Montréal, Canada: Association for Computational Linguistics, June 7–8, 2012, pp. 130–134.

Popovic, Maja, Gispert, Adrià de, Gupta, Deepa, Lambert, Patrik, Ney, Hermann, Mariño, José B., Federico, Marcello, and Banchs, Rafael E. (2006). "Morpho-syntactic information for automatic error analysis of statistical machine translation output". In: *Proceedings on the Workshop on Statistical Machine Translation*. Ed. by Philipp Koehn and Christof Monz. New York City: Association for Computational Linguistics, June 8–9, 2006, pp. 1–6.

Popović, Maja and Burchardt, Aljoscha (2011). "From human to automatic error classification for machine translation output". In: *Proceedings of the 15th International Conference of the European Association for Machine Translation (EAMT)*. Ed. by Mikel L. Forcada, Heidi Depraetere, and Vincent Vandeghinste. Leuven, Belgium: European Association for Machine Translation, May 30–31, 2011, pp. 265–272.

Popović, Maja and Ney, Hermann (2007). "Word error rates: Decomposition over pos classes and applications for error analysis". In: *Proceedings of the Second Workshop on Statistical Machine Translation*. Ed. by Chris Callison-Burch, Philipp Koehn, Christof Monz, and Cameron Shaw Fordyce. StatMT '07. Prague, Czech Republic: Association for Computational Linguistics, June 23, 2007, pp. 48–55.

Popović, Maja, Vilar, David, Avramidis, Eleftherios, and Burchardt, Aljoscha (2011). "Evaluation without references: IBM1 scores as evaluation metrics". In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Ed. by Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar F. Zaidan. Edinburgh, Scotland: Association for Computational Linguistics, July 30–31, 2011, pp. 99–103.

Przybocki, Mark, Peterson, Kay, and Bronsart, Sébastien (2009a). *2008 NIST Metrics for Machine Translation (MetricsMATR08) Development Data LDC2009T05*. [online] Available from https://catalog.ldc.upenn.edu/LDC2009T05. [Accessed on 6 July 2017]. Philadelphia.

Przybocki, Mark, Peterson, Kay, Bronsart, Sébastien, and Sanders, Gregory (2009b). "The NIST 2008 metrics for machine translation challenge—overview, methodology, metrics, and results". *Machine Translation*, 23(2) (Sept. 2009), pp. 71–103.

Q. Zadeh, Behrang and Handschuh, Siegfried (2014). "The ACL RD-TEC: A Dataset for benchmarking terminology extraction and classification in computational linguistics". In: *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*. Ed. by Patrick Drouin, Natalia Graba, Thierry Hamon, and Kyo Kageura. Dublin, Ireland: Association for Computational Linguistics and Dublin City University, Aug. 23, 2014, pp. 52–63.

Quirk, Christopher (2004). "Training a sentence-level machine translation confidence measure." In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004*. Ed. by Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, and Raquel Silva. Lisbon, Portugal: European Language Resources Association (ELRA), May 26–28, 2004, pp. 825–828.

Rabinovich, Ella and Wintner, Shuly (2015). "Unsupervised Identification of Translationese". *Transactions of the Association for Computational Linguistics*, 3 (1 2015), pp. 419–432.

Ramos, Juan et al. (2003). "Using TF-IDF to determine word relevance in document queries". In: *Proceedings of the First Instructional Conference on Machine Learning*. Ed. by Michael L. Littman, Yihua Wu, Peng Song, and Terran Lane. Piscataway, NJ, USA: Rutgers University, Dec. 3–8, 2003.

Rapp, Reinhard (1995). "Identifying word translations in non-parallel texts". In: *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*. Ed. by Hans Uszkoreit. Cambridge, Massachusetts: Association for Computational Linguistics, June 26–30, 1995, pp. 320–322.

Rasmussen, Carl Edward (2004). "Gaussian processes in machine learning". In: *Advanced Lectures on Machine Learning*. Springer, pp. 63–71.

Řehůřek, Radim and Sojka, Petr (2010). "Software framework for topic modelling with large corpora". In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Ed. by René Witte, Hamish Cunningham, Jon Patrick, Elena Beisswanger, Ekaterina Buyko, Udo Hahn, Karin Verspoor, and Anni R. Coden. http://is.muni.cz/publication/884893/en. Valletta, Malta: European Language Resources Association (ELRA), May 22, 2010, pp. 45–50.

Robertson, Stephen E and Jones, K Sparck (1976). "Relevance weighting of search terms". *Journal of the Association for Information Science and Technology*, 27(3), pp. 129–146.

Rocheteau, Jérôme and Daille, Béatrice (2011). "TTC TermSuite: A UIMA application for multilingual terminology extraction from comparable corpora". In:

Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP). Ed. by Kam-Fai Wong, Haifeng Wang, David Yarowsky, Virach Sornlertlamvanich, and Hitoshi Isahara. System Demonstrations. Chiang Mai, Thailand: Asian Federation of Natural Language Proceesing, Nov. 8–13, 2011, pp. 9–12.

Rose, Stuart, Engel, Dave, Cramer, Nick, and Cowley, Wendy (2010). "Automatic keyword extraction from individual documents". In: *Text Mining: Applications and Theory*. Ed. by Michael W. Berry and Jacob Kogan. Hoboken, New Jersey: John Wiley & Sons, Ltd, Mar. 2010. Chap. 1, pp. 1–20.

Rosenfeld, R. (2000). "Two decades of statistical language modeling: where do we go from here?" *Proceedings of the IEEE*, 88(8) (Aug. 2000), pp. 1270–1278.

Roth, Michael and Lapata, Mirella (2016). "Neural semantic role labeling with dependency path embeddings". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Katrin Erk and Noah A. Smith. Berlin, Germany: Association for Computational Linguistics, Aug. 7–12, 2016, pp. 1192–1202.

Rubino, Raphael, Foster, Jennifer, Wagner, Joachim, Roturier, Johann, Samad Zadeh Kaljahi, Rasul, and Hollowood, Fred (2012). "DCU-Symantec submission for the wmt 2012 quality estimation task". In: *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Ed. by Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Montréal, Canada: Association for Computational Linguistics, June 7–8, 2012, pp. 135–141.

Rubino, Raphael, Souza, Jose Guilherme Camargo de, Foster, Jennifer, and Specia, Lucia (2013a). "Topic models for translation quality estimation for gisting purposes". In: *Proceedings of Machine Translation Summit XIV*. Ed. by Andy Way, Khalil Sima'an, Mikel L. Forcada, Daniel Grasmick, and Heidi Depaetere. Nice,France: International Association for Machine Translation, Sept. 2–6, 2013, pp. 295–302.

Rubino, Raphael, Wagner, Joachim, Foster, Jennifer, Roturier, Johann, Samad Zadeh Kaljahi, Rasoul, and Hollowood, Fred (2013b). "DCU-Symantec at the WMT 2013 quality estimation shared task". In: *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Chris Callison-Burch, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Hervé Saint-Amand, Radu Soricut, and Lucia Specia. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 8–9, 2013, pp. 392–397.

Ruder, Sebastian (2017). "A survey of cross-lingual embedding models". *CoRR*, abs/1706.04902.

SAE (2001). *SAE J2450 translation quality metric*. Ed. by http://standards.sae.org/. [Online] Available from http://www.apex-translations.com/documents/sae_j2450.pdf. [Accessed on 25 July 2017]. Dec. 2001.

Samad Zadeh Kaljahi, Rasoul, Foster, Jennifer, Rubino, Raphael, Roturier, Johann, and Hollowood, Fred (2013). "Parser accuracy in quality estimation of machine translation: A tree kernel approach". In: *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Ed. by Ruslan Mitkov and Jong C. Park. Nagoya, Japan: Asian Federation of Natural Language Processing, Oct. 14–18, 2013, pp. 1092–1096.

Santos, Cicero Dos and Zadrozny, Bianca (2014). "Learning Character-level Representations for Part-of-Speech Tagging". In: *Proceedings of the 31st International Conference on Machine Learning*. Ed. by Eric P. Xing and Tony Jebara. Vol. 32.

Proceedings of Machine Learning Research 2. Bejing, China: PMLR, June 22–24, 2014, pp. 1818–1826.

Saralegi, Xabier, San Vicente, Iñaki, and Gurrutxaga, Antton (2008). “Automatic extraction of bilingual terms from comparable corpora in a popular science domain”. In: *Proceedings of Building and using Comparable Corpora Workshop*. Ed. by Pierre Zweigenbaum, Éric Gaussier, and Pascale Fung. Miyazaki,Japan: European Language Resources Association (ELRA), May 31, 2008, pp. 27–32.

Scarton, Carolina (2016). “Document-level machine translation quality estimation”. PhD thesis. Sheffield, UK: University of Sheffield.

Scarton, Carolina, Beck, Daniel, Shah, Kashif, Sim Smith, Karin, and Specia, Lucia (2016). “Word embeddings and discourse information for quality estimation”. In: *Proceedings of the First Conference on Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Rajan Chatterjee, Christian Federmann, Liane Guillou, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Aurelie Neveol, Mariana Neves, Pavel Pecina, Martin Popel, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Lucia Specia, Karin Verspoor, Joerg Tiedemann, and Marco Turchi. Berlin, Germany: Association for Computational Linguistics, Aug. 11–12, 2016, pp. 831–837.

Scarton, Carolina and Specia, Lucia (2014a). “Document-level translation quality estimation: exploring discourse and pseudo-references”. In: *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*. Ed. by Marko Tadić, Johann Roturier, and Philipp Koehn. EAMT. Dubrovnik, Croatia: European Association for Machine Translation, June 16–18, 2014, pp. 101–108.

Scarton, Carolina and Specia, Lucia (2014b). “Exploring consensus in machine translation for quality estimation”. In: *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Matouš Macháček, Christof Monz, Pavel Pecina, Matt Post, Hervé Saint-Amand, Radu Soricut, and Lucia Specia. Baltimore, Maryland, USA: Association for Computational Linguistics, June 26–27, 2014, pp. 342–347.

Scarton, Carolina, Tan, Liling, and Specia, Lucia (2015a). “USHEF and USAAR-USHEF participation in the WMT15 QE shared task”. In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, Pavel Pecina, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Hervé Saint-Amand, Carolina Scarton, Lucia Specia, and Marco Turchi. Lisbon, Portugal: Association for Computational Linguistics, Sept. 17–18, 2015, pp. 336–341.

Scarton, Carolina, Zampieri, Marcos, Vela, Mihaela, Genabith, Josej van, and Specia, Lucia (2015b). “Searching for context: A study on document-level labels for translation quality estimation”. In: *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*. Ed. by Iknur Durgar El-Kahlout, Mehmed Özkan, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, Fred Hollowood, and Andy Way. EAMT. Antalya, Turkey: European Association for Machine Translation (EAMT), May 11–13, 2015, pp. 121–128.

Schwartz, Roy, Thomson, Sam, and Smith, Noah A. (2018). “Bridging CNNs, RNNs, and Weighted Finite-State Machines”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Iryna Gurevych and Yusuke Miyao. Melbourne, Australia: Association for Computational Linguistics, July 15–20, 2018, pp. 295–305.

Sclano, F. and Velardi, P. (2007). “TermExtractor: A web application to learn the shared terminology of emergent web communities”. In: *Enterprise Interoperability II: New Challenges and Approaches*. Ed. by Kai Mertins Ricardo J. Gonçalves Jörg P. Müller and Martin Zelm. London: Springer, pp. 287–290.

Secară, Alina (2005). “Translation evaluation–a state of the art survey”. In: *Proceedings of the eCoLoRe/MeLLANGE Workshop*. [Online] Available from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.126.3654&rep=rep1&type=pdf. [Accessed on 28 May 2017]. Leeds, UK, pp. 39–44.

Sedgwick, Philip et al. (2012). “Pearson’s correlation coefficient”. *BMJ*, 345(7).

Seginer, Yoav (2007). “Learning syntactic structure”. PhD thesis. Amsterdam, Netherlands: Institute for Logic, Language and Computation.

Servan, Christophe, Le, Ngoc-Tien, Luong, Ngoc Quang, Lecouteux, Benjamin, and Besacier, Laurent (2015). “An open source toolkit for word-level confidence estimation in machine translation”. In: *The 12th International Workshop on Spoken Language Translation (IWSLT’15)*. Ed. by Marcello Federico, Sebastian Stüker, and Jan Niehues. Da Nang, Vietnam: Advanced Telecommunication Research Institute International (ATR), Dec. 3–4, 2015.

Sha, Lei, Li, Sujian, Chang, Baobao, Sui, Zhifang, and Jiang, Tingsong (2016). “Capturing Argument Relationship for Chinese Semantic Role Labeling”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Ed. by Xavier Carreras and Kevin Duh. Austin, Texas: Association for Computational Linguistics, Nov. 1–5, 2016, pp. 2011–2016.

Shah, Kashif, Conn, Trevor, and Specia, Lucia (2013). “An Investigation on the effectiveness of features for translation quality estimation”. In: *Proceedings of the Machine Translation Summit XIV*. Ed. by Andy Way, Khalil Sima’an, Mikel L. Forcada, Daniel Grasmick, and Heidi Depaetere. Nice, France: International Association for Machine Translation, Sept. 2–6, 2013, pp. 167–174.

Shah, Kashif, Logacheva, Varvara, Paetzold, Gustavo, Blain, Frédéric, Beck, Daniel, Bougares, Fethi, and Specia, Lucia (2015). “SHEF-NN: Translation quality estimation with neural networks”. In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, Pavel Pecina, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Hervé Saint-Amand, Carolina Scarton, Lucia Specia, and Marco Turchi. Lisbon, Portugal: Association for Computational Linguistics, Sept. 17–18, 2015, pp. 342–347.

Shang, Liugang, Cai, Dongfeng, and Ji, Duo (2015). “Strategy-based technology for estimating MT quality”. In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, Pavel Pecina, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Hervé Saint-Amand, Carolina Scarton, Lucia Specia, and Marco Turchi. Lisbon, Portugal: Association for Computational Linguistics, Sept. 17–18, 2015, pp. 348–352.

Sharoff, Serge (2018). “Language adaptation experiments via cross-lingual embeddings for related languages”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga. Miyazaki, Japan: European Language Resources Association (ELRA), May 7–12, 2018, pp. 844–849.

Smith, Karin Sim, Aziz, Wilker, and Specia, Lucia (2015). “A proposal for a coherence corpus in machine translation”. In: *Proceedings of the Second Workshop on Discourse in Machine Translation*. Ed. by Bonnie Webber, Marine Carpuat, Andrei Popescu-Belis, and Christian Hardmeier. Lisbon, Portugal: Association for Computational Linguistics, Sept. 17, 2015, pp. 52–58.

Snover, Matthew, Dorr, Bonnie, Schwartz, Richard, Micciulla, Linnea, and Makhoul, John (2006). “A study of translation edit rate with targeted human annotation”. In: *Proceedings of 7th Conference of the Association for Machine Translation in the Americas*. Ed. by Laurie Gerber, Nizar Habash, and Alon Lavie. Cambridge, Massachusetts, USA: Association for Machine Translation of the Americas, Aug. 8–12, 2006, pp. 223–231.

Socher, Richard, Huval, Brody, Manning, Christopher D., and Ng, Andrew Y. (2012). “Semantic compositionality through recursive matrix-vector spaces”. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Ed. by Junichi Tsujii, James Henderson, and Marius Paşca. EMNLP-CoNLL ’12. Jeju Island, Korea: Association for Computational Linguistics, July 12–14, 2012, pp. 1201–1211.

Socher, Richard, Pennington, Jeffrey, Huang, Eric H., Ng, Andrew Y., and Manning, Christopher D. (2011). “Semi-supervised recursive autoencoders for predicting sentiment distributions”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Ed. by Paola Merlo, Regina Barzilay, and Mark Johnson. EMNLP ’11. Edinburgh, United Kingdom: Association for Computational Linguistics, July 27–31, 2011, pp. 151–161.

Solano-Flores, Guillermo, Backhoff, Eduardo, and Contreras-Niño, Luis Ángel (2009). “Theory of test translation error”. *International Journal of Testing*, 9(2), pp. 78–91.

Soricut, Radu, Bach, Nguyen, and Wang, Ziyuan (2012). “The SDL language weaver systems in the wmt12 quality estimation shared Task”. In: *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Ed. by Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Montréal, Canada: Association for Computational Linguistics, June 7–8, 2012, pp. 142–148.

Soricut, Radu and Echihabi, Abdessamad (2010). “TrustRank: Inducing trust in automatic translations via ranking”. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Ed. by Jan Hajič, Sandra Carberry, Stephen Clark, and Joakim Nivre. Uppsala, Sweden: Association for Computational Linguistics, July 11–16, 2010, pp. 612–621.

Soricut, Radu and Narsale, Sushant (2012). “Combining quality prediction and system selection for improved automatic translation output”. In: *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Ed. by Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Montréal, Canada: Association for Computational Linguistics, June 7–8, 2012, pp. 163–170.

Spearman, Charles (1904). “The proof and measurement of association between two things”. *The American journal of psychology*, 15(1), pp. 72–101.

Specia, Lucia (2011). “Exploiting objective annotations for minimising translation post-editing effort”. In: *Proceedings of the 15th International Conference of the European Association for Machine Translation (EAMT)*. Ed. by Mikel L. Forcada, Heidi Depraetere, and Vincent Vandeghinste. Leuven, Belgium: European Association for Machine Translation, May 30–31, 2011, pp. 73–80.

Specia, Lucia, Cancedda, Nicola, Dymetman, Marc, Turchi, Marco, and Cristianini, Nello (2009a). “Estimating the sentence-level quality of machine translation

systems". In: *Proceedings of the 13th Conference of the European Association for Machine Translation*. Ed. by Lluís Márquez and Harold Somers. Barcelona, Spain: European Association for Machine Translation (EAMT), May 14–15, 2009, pp. 28–37.

Specia, Lucia, Hajlaoui, Najeh, Hallett, Catalina, and Aziz, Wilker (2011). "Predicting machine translation adequacy". In: *Proceedings of the 13th Machine Translation Summit (MT Summit XIII)*. Ed. by Le Sun, Xiaodong Shi, and Qun Liu. Xiamen, China: International Association for Machine Translation, Sept. 19–23, 2011, pp. 513–520.

Specia, Lucia and Logacheva, Varvara (2017). *WMT17 quality estimation shared task training and development data*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Specia, Lucia, Logacheva, Varvara, Blain, Frederic, Fernandez, Ramon, and Martins, André (2018). *WMT18 quality estimation shared task training and development data*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Specia, Lucia, Paetzold, Gustavo, and Scarton, Carolina (2015). "Multi-level translation quality prediction with QuEst++". In: *Proceedings of The 53rd Annual Meeting of the Association for Computational Linguistics and The 7th International Joint Conference on Natural Language Processing: System Demonstrations*. Ed. by Hsin-Hsi Chen and Katja Markert. Beijing, China: Association for Computational Linguistics, July 26–31, 2015, pp. 115–120.

Specia, Lucia, Raj, Dhwaj, and Turchi, Marco (2010). "Machine translation evaluation versus quality estimation". *Machine Translation*, 24(1), pp. 39–50.

Specia, Lucia, Shah, Kashif, Souza, Jose G.C. de, and Cohn, Trevor (2013). "QuEst - A translation quality estimation framework". In: *Proceedings of The 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Ed. by Miriam Butt and Sarmad Hussain. ACL. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 4–9, 2013, pp. 79–84.

Specia, Lucia, Turchi, Marco, Wang, Zhuoran, Shawe-Taylor, John, and Saunders, Craig (2009b). "Improving the confidence of machine translation quality estimates". In: *Proceedings of the Twelfth Machine Translation Summit (MT Summit XII)*. Ed. by Pierre Isabelle, Roland Kuhn, Nick Bemish, Mike Dillinger, and Marie-Josée Goulet. Ottawa, Ontario, Canada: International Association for Machine Translation, Aug. 26–30, 2009.

Srivastava, Nitish, Hinton, Geoffrey E, Krizhevsky, Alex, Sutskever, Ilya, and Salakhutdinov, Ruslan (2014). "Dropout: a simple way to prevent neural networks from overfitting." *Journal of machine learning research*, 15(1), pp. 1929–1958.

Stanojević, Miloš and Sima'an, Khalil (2014). "Fitting sentence level translation evaluation with many dense features". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Alessandro Moschitti, Bo Pang, and Walter Daelemans. Doha, Qatar: Association for Computational Linguistics, Oct. 25–29, 2014, pp. 202–206.

Steele, David (2015). "Improving the translation of discourse markers for Chinese into English". In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. Ed. by Diana Inkpen, Smaranda Muresan, Shibamouli Lahiri, Karen Mazidi, and Alisa Zhila. Denver, Colorado: Association for Computational Linguistics, June 1, 2015, pp. 110–117.

Strubell, Emma, Verga, Patrick, Belanger, David, and McCallum, Andrew (2017). "Fast and Accurate Entity Recognition with Iterated Dilated Convolutions". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Ed. by Martha Palmer, Rebecca Hwa, and Sebastian Riedel. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 9–11, 2017, pp. 2670–2680.

Stymne, Sara and Ahrenberg, Lars (2012). "On the practice of error analysis for machine translation evaluation". In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Istanbul, Turkey: European Language Resources Association (ELRA), May 23–25, 2012, pp. 1785–1790.

Tai, Kai Sheng, Socher, Richard, and Manning, Christopher D. (2015). "Improved semantic representations from tree-structured Long Short-Term Memory neetworks". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Chengqing Zong and Michael Strube. Beijing, China: Association for Computational Linguistics, July 25–31, 2015, pp. 1556–1566.

Tang, Duyu, Qin, Bing, and Liu, Ting (2015). "Document Modeling with Gated Recurrent Neural Network for Sentiment Classification". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Ed. by Chris Callison-Burch and Jian Su. Lisbon, Portugal: Association for Computational Linguistics, Sept. 17–21, 2015, pp. 1422–1432.

Tang, Duyu, Wei, Furu, Yang, Nan, Zhou, Ming, Liu, Ting, and Qin, Bing (2014). "Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Kristina Toutanova and Hua Wu. Baltimore, Maryland: Association for Computational Linguistics, June 22–27, 2014, pp. 1555–1565.

Tezcan, Arda, Hoste, Veronique, Desmet, Bart, and Macken, Lieve (2015). "UGENT-LT3 SCATE system for machine translation quality estimation". In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, Pavel Pecina, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Hervé Saint-Amand, Carolina Scarton, Lucia Specia, and Marco Turchi. Lisbon, Portugal: Association for Computational Linguistics, Sept. 17–18, 2015, pp. 353–360.

Tezcan, Arda, Hoste, Véronique, and Macken, Lieve (2016). "UGENT-LT3 SCATE submission for WMT16 shared task on quality estimation". In: *Proceedings of the First Conference on Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Rajan Chatterjee, Christian Federmann, Liane Guillou, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Aurelie Neveol, Mariana Neves, Pavel Pecina, Martin Popel, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Lucia Specia, Karin Verspoor, Joerg Tiedemann, and Marco Turchi. Berlin, Germany: Association for Computational Linguistics, Aug. 11–12, 2016, pp. 843–850.

Thomas, Jenny and Short, Mick (1996). *Using corpora for language research: Studies in Honour of Geoffrey Leech*. London: Longman.

Tian, Liang, Wong, Derek F., Chao, Lidia S., Quaresma, Paulo, Oliveira, Francisco, and Yi, Lu (2014). "UM-corpus: A large English-Chinese parallel corpus for statistical machine translation". In: *Proceedings of the Ninth International Conference*

on Language Resources and Evaluation (LREC'14). Ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Reykjavik, Iceland: European Language Resources Association (ELRA), May 26–31, 2014, pp. 1837–1842.

Tiedemann, Jörg (2011). "Bitext alignment". *Synthesis Lectures on Human Language Technologies*, 4(2), pp. 1–165.

Tieleman, Tijmen and Hinton, Geoffrey (2012). *Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. in COURSERA: Neural Networks for Machine Learning*. Tech. rep. University of Toronto.

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., and Sawaf, H. (1997). "Accelerated dp based search for statistical translation". In: *Proceedings of European Conference on Speech Communication and Technology*. Ed. by George Kokkinakis, Nikos Fakotakis, and Evangelos Dermatas. Rhodes, Greece, Sept. 22–25, 1997, pp. 2667–2670.

Toury, Gideon (2012). *Descriptive translation studies and beyond: revised edition*. Vol. 100. John Benjamins Publishing.

Turchi, Marco, Negri, Matteo, and Federico, Marcello (2013). "Coping with the subjectivity of human judgements in MT quality estimation". In: *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Chris Callison-Burch, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Hervé Saint-Amand, Radu Soricut, and Lucia Specia. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 8–9, 2013, pp. 240–251.

Ueffing, Nicola, Macherey, Klaus, and Ney, Hermann (2003). "Confidence measures for statistical machine translation". In: *Proceedings of the Ninth Machine Translation Summit*. Ed. by Eduard Hovy and Elliott Macklovitch. New Orleans, USA: Springer-Verlag, Sept. 23–27, 2003, pp. 394–401.

Ueffing, Nicola and Ney, Hermann (2005). "Word-Level confidence estimation for machine translation using phrase-based translation models". In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Ed. by Chris Brew, Lee-Feng Chien, and Katrin Kirchhoff. Vancouver, British Columbia, Canada: Association for Computational Linguistics, Oct. 6–8, 2005, pp. 763–770.

Ueffing, Nicola and Ney, Hermann (2007). "Word-Level confidence estimation for machine translation". *Computational Linguistics*, 33(1), pp. 9–40.

Vapnik, Vladimir N. (1998). *Statistical learning theory*. New York: John Wiley & Sons, Ltd.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Łukasz, and Polosukhin, Illia (2017a). "Attention is All you Need". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Long Beach, California, USA: Curran Associates, Inc., Dec. 4–9, 2017, pp. 5998–6008.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Łukasz, and Polosukhin, Illia (2017b). "Attention is all you need". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., pp. 5998–6008.

Veal, L Ramon and Hudson, Sally Ann (1983). “Direct and indirect measures for large-scale evaluation of writing”. *Research in the Teaching of English*, 17(3), pp. 290–296.

Vilar, David, Xu, Jia, d’Haro, Luis Fernando, and Ney, Hermann (2006). “Error analysis of statistical machine translation output”. In: *Proceedings of The Fifth International Conference on Language Resources and Evaluation*. Ed. by Nicoletta Calzolari, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk, and Daniel Tapias. Genoa, Italy: European Language Resources Association (ELRA), May 24–26, 2006, pp. 697–702.

Vinay, J. -. and Darbelnet, J. (1958). *Stylistique comparée du français et de l’anglais: méthode de traduction*. Vol. 1. London;Paris; Didier.

Vintar, Spela (2010). “Bilingual term recognition revisited: The bag-of-equivalents term alignment approach and its evaluation”. *Terminology : International Journal of Theoretical and Applied Issues in Specialized Communication*, 16(2), pp. 141–158.

Voutilainen, Atro (2003). “Part-of-speeching tagging”. In: *The Oxford Handbook of Computational Linguistics*. Ed. by Ruslan Mitkov. Oxford: Oxford University Press. Chap. 11, pp. 219–233.

Wan, Xiaojun and Peng, Yuxin (2005). “The earth mover’s distance as a semantic measure for document similarity”. In: *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*. Ed. by Abudur Chowdhury, Norbert Fuhr, Marc Ronthaler, Hans-Jörg, and Wilfried Teiken. Bremen, Germany: Association for Computing Machinery, Oct. 31–Nov. 5, 2005, pp. 301–302.

Warburton, Kara (2013). “Processing terminology for the translation pipeline”. *Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication*, 19(1), p. 93.

Webber, Bonnie, Carpuat, Marine, Popescu-Belis, Andrei, and Hardmeier, Christian, eds. (2015). *Proceedings of the Second Workshop on Discourse in Machine Translation*. Lisbon, Portugal: Association for Computational Linguistics, Sept. 17, 2015.

Webber, Bonnie, Popescu-Belis, Andrei, Markert, Katja, and Tiedemann, Jörg, eds. (2013). *Proceedings of the Workshop on Discourse in Machine Translation*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 9, 2013.

Webber, Bonnie, Popescu-Belis, Andrei, and Tiedemann, JÃ¶rg, eds. (2017). *Proceedings of the Third Workshop on Discourse in Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 8, 2017.

Weigel, A. and Fein, F. (1994). “Normalizing the weighted edit distance”. In: *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3 - Conference C: Signal Processing (Cat. No.94CH3440-5)*. Ed. by S. Peleg and S. Ullman. Vol. 2. Jerusalem, Israel, Israel: IEEE, Oct. 9–13, 1994, pp. 399–402.

Wen, Qiufang and Wang, Jinquan (2008). *Parallel Corpus of Chinese EFL Learners*. Beijing,China: Foreign Language Teaching and Research Press.

Wermter, Joachim and Hahn, Udo (2005). “Finding new terminology in very large corpora”. In: *Proceedings of the 3rd International Conference on Knowledge Capture*. Ed. by Peter Clark and Guus Schreiber. Banff, Alberta, Canada: Association for Computing Machinery, Oct. 2–5, 2005, pp. 137–144.

White, John S. (1994). “The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Further Approaches”. In: *Proceedings of the First Conference of the As-*

sociation for Machine Translation in the Americas. Columbia, Maryland, USA: Association for Machine Translation in the Americas, Oct. 5–8, 1994, pp. 193–205.

Wisniewski, Guillaume, Pécheux, Nicolas, Allauzen, Alexander, and Yvon, François (2014). "LIMSI submission for WMT'14 QE task". In: *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Matouš Macháček, Christof Monz, Pavel Pecina, Matt Post, Hervé Saint-Amand, Radu Soricut, and Lucia Specia. Baltimore, Maryland, USA: Association for Computational Linguistics, June 26–27, 2014, pp. 348–354.

Wong, Billy T. M. and Kit, Chunyu (2012). "Extending machine translation evaluation metrics with lexical cohesion to document level". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Ed. by Junichi Tsujii, James Henderson, and Marius Paşca. Jeju Island, Korea: Association for Computational Linguistics, July 12–14, 2012, pp. 1060–1068.

Writh, Jarryl and Catlett, Jason (1988). "Experiments on the costs and benefits of windowing in {ID3}". In: *Machine Learning Proceedings 1988*. Ed. by John Laird. 1st ed. San Francisco, California: Morgan Kaufmann, Dec. 25, 1988, pp. 87–99.

Wu, Chunyang and Zhao, Hai (2012). "Regression with phrase indicators for estimating MT quality". In: *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Ed. by Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Montréal, Canada: Association for Computational Linguistics, June 7–8, 2012, pp. 149–153.

Wu, Dekai (1997). "Stochastic inversion transduction grammars and bilingual parsing of parallel corpora". *Computational Linguistics*, 23(3) (Sept. 1997), pp. 377–403.

Wu, Dekai (2010). "Alignment". In: *Handbook of Natural Language Processing, Second Edition*. Ed. by Nitin Indurkhya and Fred J. Damerau. Boca Raton, Florida: CRC Press, Taylor and Francis Group. Chap. 16, pp. 367–408.

Wu, Haiyang, Dong, Daxiang, Hu, Xiaoguang, Yu, Dianhai, He, Wei, Wu, Hua, Wang, Haifeng, and Liu, Ting (2014). "Improve statistical machine translation with context-sensitive bilingual semantic embedding model". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Alessandro Moschitti, Bo Pang, and Walter Daelemans. Doha, Qatar: Association for Computational Linguistics, Oct. 25–29, 2014, pp. 142–146.

Xia, Tian and Chai, Yanmei (2011). "An improvement to TF-IDF: Term distribution based term weight algorithm". *Jouranl of Software*, 6(3) (Mar. 2011), pp. 413–420.

Xiong, D., Zhang, M., and Wang, X. (2015). "Topic-based coherence modeling for statistical machine translation". *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3) (Mar. 2015), pp. 483–493.

Xiong, Deyi and Zhang, Min (2013). "A topic-based coherence model for statistical machine translation". In: *Proceedings of The Twenty-Seventh AAAI Conference on Artificial Intelligence*. Ed. by Marie des Jardins, Michael Littman, Hector Munoz-Avila, David Stracuzzi, Laura E. Brown, and David Kauchak. Bellevue, Washington, July 14–18, 2013, pp. 977–983.

Xu, Ran and Sharoff, Serge (2014). "Evaluating term extraction methods for interpreters". In: *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*. Ed. by Patrick Drouin, Natalia Grabar, Thierry Hamon, and Kyo Kageura. Dublin, Ireland: Association for Computational Linguistics and Dublin City University, Aug. 23, 2014, pp. 86–93.

Xu, Rui and Wunsch, D. (2005). “Survey of clustering algorithms”. *IEEE Transactions on Neural Networks*, 16(3) (May 2005), pp. 645–678.

Xu, Yan, Chen, Luoxin, Wei, Junsheng, Ananiadou, Sophia, Fan, Yubo, Qian, Yi, Eric, I, Chang, Chao, and Tsujii, Junichi (2015). “Bilingual term alignment from comparable corpora in English discharge summary and Chinese discharge summary”. *BMC bioinformatics*, 16(1).

Yang, Yiming and Pedersen, Jan O. (1997). “A comparative study on feature selection in text categorization”. In: *Proceedings of the Fourteenth International Conference on Machine Learning (ICML 1997)*. Ed. by Douglas H. Fisher. Nashville, Tennessee, USA, Morgan Kaufmann Publishers Inc., July 8–12, 1997, pp. 412–420.

Yin, Wenpeng, Schütze, Hinrich, Xiang, Bing, and Zhou, Bowen (2016). “ABCNN: Attention-based convolutional neural network for modeling sentence pairs”. *Transactions of the Association for Computational Linguistics*, 4 (1 2016), pp. 259–272.

Yu, Lei, Hermann, Karl Moritz, Blunsom12, Phil, and Pulman, Stephen (2014). “Deep Learning for answer sentence selection”. In: *Deep Learning and Representation Learning Workshop: NIPS 2014*. Ed. by Yoshua Bengio, Adam Coates, Roland Memisevic, Andrew Ng, and Daan Wierstra. Montreal, Canada., Dec. 12, 2014.

Yuan, Yu (2016). “A Feature Set for automatted human translation quality estimation”. *Foreign Language Teaching and Research*, 49(5), pp. 776–787.

Yuan, Yu, Sharoff, Serge, and Babych, Bogdan (2016). “MoBiL: A hybrid feature set for automatic human translation quality assessment”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Portorož, Slovenia: European Language Resources Association (ELRA), May 23–28, 2016, pp. 3663–3670.

Zhang, Bo, Xiao, Yunnan, and Luo, Juan (2015). “Rater reliability and score discrepancy under holistic and analytic scoring of second language writing”. *Language Testing in Asia*, 5(1), pp. 54–62.

Zhang, X., Song, Y., and Fang, A. C. (2010). “Term recognition using conditional random fields”. In: *Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering (NLPKE-2010)*. Ed. by Fuji Ren and Yixin Zhong. Beijing,China: IEEE Xplore Digital Library, Aug. 21–23, 2010, pp. 1–6.

Zhang, Ying and Vogel, Stephan (2005). “An efficient phrase-to-phrase alignment model for arbitrarily long phrase and large corpora”. In: *Proceedings of the 10th EAMT Conference Practical Application of Machine Translation*. Ed. by Bente Maegaard, Viggo Hansen, Steven Krauwer, Gábor Prószéky, Harold Somers, and Gregor Thurmair. Budapest, Hungary: European Association for Machine Translation, May 30–31, 2005, pp. 294–301.

Zhang, Ying, Vogel, Stephan, and Waibel, Alex (2004). “Interpreting bleu/nist scores: How much improvement do we need to have a better system?” In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Teresa Lino, Bente Maegaard, Joseph Mariani, Jan Odijk, and Daniel Tapias. Lisbon, Portugal: European Language Resources Association (ELRA), May 26–28, 2004, pp. 2051–2054.

Zhang, Yue, Liu, Qi, and Song, Linfeng (2018). “Sentence-State LSTM for Text Representation”. In: *Proceedings of the 56th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers). Ed. by Iryna Gurevych and Yusuke Miyao. Melbourne, Australia: Association for Computational Linguistics, July 15–20, 2018, pp. 317–327.

Zhang, Ziqi, Gao, Jie, and Ciravegna, Fabio (2016). "JATE 2.0: Java automatic term extraction with Apache Solr". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Portorož, Slovenia: European Language Resources Association (ELRA), May 23–28, 2016, pp. 2262–2269.

Zhang, Ziqi, Iria, Jose, Brewster, Christopher, and Ciravegna, Fabio (2008). "A comparative evaluation of term recognition algorithms". In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias. Marrakech, Morocco: European Language Resources Association (ELRA), May 28, 2008.

Zhao, J., Lan, M., Niu, Z. Y., and Lu, Yue (2015). "Integrating word embeddings and traditional NLP features to measure textual entailment and semantic relatedness of sentence pairs". In: *2015 International Joint Conference on Neural Networks (IJCNN)*. Ed. by De-shuang Huang, Yoonsuck Choe, Haibo He, and Asim Roy. Killarney, Ireland: IEEE Xplore Digital Library, July 12–17, 2015, pp. 1–7.

Zhou, GuoDong and Su, Jian (2004). "Exploring deep knowledge resources in biomedical name recognition". In: *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications*. Ed. by Nigel Collier, Patrick Ruch, and Adeline Nazarenko. JNLPBA '04. Geneva, Switzerland: Association for Computational Linguistics, Aug. 28–29, 2004, pp. 96–99.

Zou, Will Y., Socher, Richard, Cer, Daniel, and Manning, Christopher D. (2013). "Bilingual word embeddings for phrase-based machine translation". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Ed. by David Yarowsky, Timothy Baldwin, and Anna Korhonen. Seattle, Washington, USA: Association for Computational Linguistics, Oct. 18–21, 2013, pp. 1393–1398.

Appendix A

Source Texts and Examples of Trainee Translations

This appendix contains source texts in English and one example trainee translations to them each, together with a Google Translate back translation for comparison.

Source Text	Trainee Translation Example	Back Translation
hostility and fear are universally the products of ignorance , and our antagonism is no exception . One cannot deny that insects are a nuisance when their bites become sore , and a threat when they transmit disease , but , viewed dispassionately , even noxious insects are beautiful . From having studied their fossilized remains , we know that insects have inhabited the earth for nearly 400 million years . Today we find them abundantly everywhere we look . Insects have been discovered on the snow of the polar caps and the peaks of high mountains . Bees , wasps , ants and termites have intricate societies in which different members are specialized for foraging , defense and reproduction . The life of a worker honeybee is even separated into successive occupations : during the first three weeks the young worker grooms the queen and her eggs , cleans out the hive , cools it by wing-fanning at the entrance , and attacks or walls in intruders . Only after this apprenticeship is the graduate allowed to leave the hive and forage for nectar and pollen . Add to such behavior the fact that some ants use leaf fragments as spoons in which to carry soft food back to their nest , and one is tempted to describe insects as " intelligent " and begin to make comparisons between insect and human societies . It saddens me that people deny themselves the simple pleasure of appreciating the exquisite elegance of even the more common insects . Only a moment with a hand lens is required to discover a new world of design and beauty .	敌意和恐惧常常被认为是无知的产物， 我们对昆虫的敌对情绪也不例外。没有人可以否认昆虫是一个令人讨厌的家伙， 因为她咬我们会给我们带来疼痛并会传播疾病。但是， 客观地说， 其实有毒的昆虫是美丽的。从对它们遗留下来的化石的研究来看， 我们了解到昆虫在地球上居住了将近400万年。今天我们所见之处都有它们的身影。在冰雪覆盖的极地地区和高山的山峰也发现了昆虫的踪迹。蜜蜂， 黄蜂， 蚂蚁和白蚁杂居在一起， 形成了特殊的社会群体， 各成员分工不同有的负责寻食， 有的负责防御还有的负责繁殖。工蜂的一生甚至被分成连续的工作阶段： 在最初的三周里年轻的工蜂为母蜂和幼蜂擦洗， 打扫蜂窝， 在蜂窝口用翅膀扇风为蜂窝降温并防御外敌入侵。学徒期结束便进入下一阶段工作， 为蜂窝产生花蜜和传播花粉做好准备。除此之外， 一些蚂蚁用碎树叶子当作匙子一样将软食物输送回巢穴。有人认为昆虫很聪明并开始拿它和人类作比较。令我难过的是人们拒绝接受欣赏极其优雅的东西带来的简单快乐， 甚至对极其普通的昆虫也是如此。发现一个崭新世界的创意和美丽， 用一只隐形的手， 那只是片刻的事情。	Hostility and fear are often considered to be the product of ignorance, and our hostility to insects is no exception. No one can deny that an insect is an annoying person because she bites us and causes pain and spreads disease. However, objectively speaking, actually poisonous insects are beautiful. From the study of the fossils that they left behind, we understand that insects have lived on Earth for nearly 4 million years. Today we have their presence everywhere we see. Traces of insects have also been found on snow-covered polar regions and mountain peaks. Bees, wasps, ants and termites are living together to form a special social group. Each member has a different division of labor and is responsible for searching for food, and some are responsible for defense and others responsible for breeding. The life of a worker bee is even divided into successive stages of work: During the first three weeks, the young worker bees scrub the mother and baby bees, sweep the honeycomb, use the wings of the honeycomb to cool the honeycomb and defend against foreign invasion. When the apprentice period is over, he enters the next stage of work, preparing the nectar for the honeycomb and spreading the pollen. In addition to this, some ants send soft foods back to their nests using broken tree leaves as a spoon. Some people think that insects are very smart and start to compare it with humans. What makes me sad is that people refuse to accept the simple happiness brought about by the extremely elegant things, even to the extremely ordinary insects. Finding a new world of creativity and beauty, using an invisible hand, is just a moment.

Table A.1 ST and Example Trainee Translation - Insects

Source Text	Trainee Translation Example	Back Translation
three types of empty-shell marriages have been identified . in a devitalized relationship husband and wife lack excitement or any real interest in their spouse or their marriage . boredom and apathy characterize this marriage . yet serious arguments are rare . in a conflict-habituated relationship husband and wife frequently quarrel in private . they may also quarrel in public or put up a facade of being compatible . the relationship is characterized by constant conflict , tension , and bitterness . and in a passive-congenial relationship both partners are not happy , but are content with their lives and generally feel adequate . the partners may have some interests in common , but these interests are generally insignificant . this type of relationship generally has little overt conflict . the number of empty-shell marriages is unknown – it may be as high as the number of happily married couples . the atmosphere in empty-shell marriages is without much fun or laughter . members do not share and discuss their problems or experiences with each other . communication is kept to a minimum . children in such families are usually starved for love and reluctant to have friends over as they are embarrassed about having their friends see their parents interacting . the couples in these marriages engage in few activities together and display no pleasure in being in one another 's company . the members are highly aware of each other 's weaknesses and sensitive areas , and they manage to frequently mention these areas in order to hurt one another .	三种 “ 空壳 ” 婚姻 已经 被 定义 了 。 在 这种 没有 什么 实质 意义 的 关系 中 ， 夫妻 对 他们 的 配偶 和 婚姻 缺乏 激情 或 真正 的 兴趣 。 厌倦 和 冷漠 是 这种 婚姻 的 典型 特征 。 尽管 激烈 的 争吵 少有 ， 但 在 这种 少闹 已成 习惯 的 关系 中 ， 双方 常 为了 自身 的 利益 而 争吵 。 他们 常常 在 表面上 保持 和谐 ， 却 也 时常 在 公共场合 争吵 。 这种 关系 充满 了 持续 的 矛盾 ， 紧张 与 苦痛 。 在 这种 消极 的 关系 中 ， 双方 都 感到 不 快乐 ， 但是 对 他们 各自 的 生活 却 普遍 觉得 比较 满足 。 双方 的 兴趣 中 可能 存在 着 某些 共性 ， 但 在 此 也 显得 无足轻重 。 这种 关系 普遍 表现 为 没什么 公开 的 矛盾 。 具体 有 多少 这种 “ 空壳 ” 婚姻 还 上 未知 的 ， 通常 它 的 数量 和 幸福 婚姻 的 一样 多 。 这种 “ 空壳 ” 婚姻 缺乏 兴趣 和 快乐 。 他们 不会 分享 和 讨论 之间 的 问题 和 经验 ， 交流 是 他们 最 缺乏 的 。 这种 家庭 的 孩子 极度 渴望 爱 ， 害怕 交朋友 ， 因为 他们 怕 父母 的 情况 让 别人 开 出来 。 这些 夫妻 很少 一起 参加 活动 ， 也 不会 因为 参与 到 对方 朋友 圈子 感到 开心 。 夫妻 双方 是 最 了解 对方 弱点 的 ， 他们 常常 利用 这些 互相 伤害 。	Three "empty shells" marriages have been defined. In this kind of relationship with no substantial meaning, couples lack passion or real interest in their spouses and marriage. Tiredness and indifference are typical features of this marriage. Although fierce quarrels are rare, in such a habit of less trouble, the two parties often quarrel over their own interests. They often seem to be in harmony, but they also often quarrel in public. This relationship is full of persistent conflicts, tensions and pains. In this negative relationship, both parties feel unhappy, but they generally feel more satisfied with their respective lives. There may be some commonalities in the interests of both parties, but this also seems to be insignificant. This kind of relationship generally shows no open contradiction. Specifically, how many such “empty shells” marriages are still unknown, usually the number of which is as much as that of happiness and marriage. This "empty shell" marriage lacks interest and happiness. They will not share the problems and experiences between discussions and discussions, and communication is the one they lack most. The children of this family are extremely eager to love and afraid of making friends because they fear that their parents' circumstances make others come forward. These couples seldom attend the event together, and they do not feel happy because they are involved in each other's friends. Both husband and wife are the ones most aware of the other's weaknesses, and they often use these to harm each other.

Table A.2 ST and Example Trainee Translation - Marriage

Source Text	Trainee Translation Example	Back Translation
i first took up walking as a means of escape . After a busy morning in my office , I found it refreshing to take a stroll at lunchtime , to breathe the fresh air and feel the sun . Another walk in the cold night air was , I discovered , an exhilarating way to unwind . I 'll never forget the feeling I got one winter night as I walked the deserted streets after many grueling hours at the hospital . I suddenly realized that I no longer felt tense or tired . All the worries about my patients¡¯ illnesses , as well as my own personal cares , seemed to evaporate as quickly as the smoky vapor of my breath in the frosty night . As I incorporated walking into my schedule , not only were my spirits lifted , but my weight and blood pressure were gradually reduced . I began reviewing the medical literature on walking . From this research , and my clinical observations as a family physician , I found that it is possible to walk your way to better health , a trimmer body and a longer life no matter what your age . Walking , like swimming , bicycling and running , is an aerobic exercise which builds the capacity for energy output and physical endurance by increasing the supply of oxygen to skin and muscles . Such exercise may be a primary factor in the prevention of heart and circulatory disease . As probably the least strenuous , safest aerobic activity , walking is the most acceptable exercise for the greatest number of people . Walking at comfortable speed improves the efficiency of the cardiorespiratory system by stimulating the lungs and heart , but at a more gradual rate than most other forms of exercise .	我 一 开始 把 散步 当作 是 一种 逃避 的 手段 。 经过 在 办公室 的 一个 忙碌 的 早晨 ， 我 发现 在 午饭时间 溜达 一下 ， 呼吸 一下 新鲜 的 空气 和 感受一下 阳光 可以 重新 使 精神 振作 。 我 发现 ， 又 一个 是 在 凉爽 夜晚 空气 中 的 散步 ， 这是 一个 令人 高兴 的 松弛 方法 。 我 不会 忘记 那个 冬天 的 夜晚 当 我 在 医院 经历 了 一大 段 使 人 精疲力竭 的 工作 时间 之后 ， 行走 在 没有 人 的 街道 上 的 感受 。 我 突然 意识 到 我 不再 感到 紧张 或 疲惫 。 我 所有 的 忧愁 ， 关于 我 的 病人 的 ， 关于 我 自己 的 ， 都 在 这个 寒冷 的 夜里 ， 随着 我 呼出 的 气息 一样 迅速 烟消云散 了 。 当 我 结合 了 散步 到 我 的 日程安排表中 ， 不仅仅 振奋 了 我 的 精神 ， 而且 我 的 体重 和 血压 也 逐步 地 降低 了 。 我 开始 再次 复习 关于 散步 的 医学 文献 。 从 这个 研究 和 我 作为 一个 家庭医生 的 临床 观测 结果 ， 我 发现 对于 任何 年龄 层次 的 人 ， 散步 对 畅通 的 呼吸 ， 更具 适应性 的 体魄 ， 和 延长 寿命 很 有 好处 。 散步 ， 就 像 游泳 ， 骑 自行车 和 跑步 一样 ， 是 一种 有氧 运动 。 它 通过 给 皮肤 和 肌肉 增加 氧气 的 提供 量 来 建立 能量 输出 的 能力 以及 增强 体力 。 这样 的 运动 在 防止 心脏 和 血液 循环系统 的 疾病 方面 起到 主要 的 作用 。 由于 散步 可能 是 最 轻松 最 安全 的 有氧 活动 ， 所以 它 是 被 最 多数 人们 接受 的 运动 。 以 舒适 的 速度 散步 能够 刺激 肺部 和 心脏 从而 提高 心肺 体系 的 功能 ， 但是 相比 其他 形式 的 运动 又 有 一个 较为 渐进 的 速率 。	In the beginning, I used walking as a means of escape. After a busy morning in the office, I found that walking around during lunch time, breathing fresh air and feeling the sun can refresh my spirits. I found that another one is taking a walk in the cool night air. This is a pleasing way to relax. I will not forget that winter night when I was in the hospital after a period of exhausting working hours and walking on the streets without people. I suddenly realized that I was no longer nervous or tired. All my worries, my patients, my own, were on this cold night, and as soon as I exhaled the breath I quickly disappeared. When I took a walk to my schedule, it not only inspired my spirit, but my weight and blood pressure also gradually decreased. I began to review the medical literature on walking again. From this study and my clinical observations as a family doctor, I found that for people of any age, walking is good for smooth breathing, more physical fitness, and longer life expectancy. Walking, like swimming, cycling and running, is an aerobic exercise. It builds energy output and enhances physical strength by increasing oxygen supply to the skin and muscles. This kind of exercise plays a major role in preventing heart and blood circulation diseases. Since walking may be the easiest and safest aerobic activity, it is the most accepted sport for most people. Walking at a comfortable pace can stimulate the lungs and heart to improve the function of the cardiopulmonary system, but there is a more gradual rate than other forms of exercise.

Table A.3 ST and Example Trainee Translation - Walking

Source Text	Trainee Translation Example	Back Translation
it was a bleak , rainy day , and I had no desire to drive up the winding mountain road to my daughter Carolyn 's house . But she had insisted that I come see something at the top of the mountain . Turning down a narrow track , we parked the car and got out . We walked along a path that was thick with old pine needles . Huge black green evergreens towered over . Gradually the peace and silence of the place began to fill my mind . Then we turned a corner , and I stopped and gasped in amazement . From the top of the mountain , sloping for several acres across folds and valleys were rivers of daffodils in radiant bloom . A profusion of color , from the palest ivory to the deepest lemon to the most vivid salmon , blazed like a carpet before us . It looked as though the sun had tipped over and spilled gold down the mountainside . A riot of questions filled my mind . Who created such beauty ? Why ? How ? As we approached the home that stood in the center of the property , we saw a sign : ANSWERS TO THE QUESTIONS I KNOW YOU ARE ASKING . The first answer was : ONE WOMAN , TWO HANDS , TWO FEET AND VERY LITTLE BRAIN . The second was : ONE AT A TIME . The third : STARTED IN 1958 . As we drove home , I was so moved by what we had seen I could scarcely speak . "She changed the world , " I finally said , "one bulb at a time " . She started almost 40 years ago , probably just the beginning of an idea , but she kept at it . " The wonder of it would not let me go . "Imagine , " I said , "if I 'd had a vision and worked at it , jut a little bit every day , what might have I accomplished ? "	这是一个灰暗的下雨天，我丝毫也不过刮着大风的山路去我女儿卡洛琳的家。但是她坚持让我去看山顶上的某样东西。开过一个狭窄的拐弯处，我们停下了车然后走出去。我们沿着一条厚厚地覆盖着古老的松针的小道走着。大片的黑绿的常绿树木呈现在眼前。渐渐的，这里的平静和安寂开始充满了我的思绪。然后我们拐了一个弯突然我仍停了下来，被这个神奇的景象吸引住了。从山顶往下看，斜跨过折层和山谷的几英亩地上都是闪着光的水仙花的海洋。一种充沛的颜色从最纯的象牙白到最深的柠檬黄再到最生动的粉红色就像一条地毯在我们面前闪耀。它看上去就像是太阳已经斜翻了过去，把它的金色倾倒在山边一样。一大串问题填满了我的脑袋。谁造就了这样的美景？为什么？怎样造就的？当我们靠近在山中的家的时候，我们看到一个标记：我知道你们在问问题的答案。第一个回答是：一个女人两只手，两只脚和很小的智慧。第二个是：一次一个。第三个是：从1958年开始。当我们开车回家的时候，我是如此得感动于我们的所见以至于我无法言语。"她改变了世界"我最后说道，"一次一个植物"。她几乎是从难从40年前开始的，可能只是一个主意的开始，但是她坚持了下来。"这个奇迹让我不能离开。"想像，"我说道，"如果我有一个想像，然后在此努力，每天只是一点，我可能会有什么成就呢？"	This is a grey rainy day, but I did not even scratch the windy mountain road to my daughter Caroline's home. But she insisted on letting me go to see something on the top of the mountain. After a narrow turn, we stopped the car and walked out. We walk along a trail thickly covered with ancient pine needles. Large patches of dark green evergreen trees appear before the eyes. Gradually, the calm and quiet here began to fill my mind. Then we turned a corner and suddenly I stopped and was attracted by this magical sight. Looking down from the top of the mountain, there are shimmering seas of daffodils on several acres of land diagonally across the fold and the valley. A plentiful color from the purest ivory to the deepest lemon yellow to the most vivid pink is like a carpet shining in front of us. It looks as though the sun has rolled over and it dumped its golden color on the edge of the hill. A large number of questions filled my head. Who made this beautiful scene? Why? How to make it? When we are close to the home in the mountains, we see a mark: I know you are asking the answer to the question. The first answer is: A woman has two hands, two feet and a small amount of wisdom. The second is：one at a time. The third is : Beginning in 1958. When we drove home, I was so touched by what we saw that I could not speak. "She changed the world." I finally said, "One plant at a time." She hardly started from 40 years ago and may just be the beginning of an idea, but she insists on it. "This miracle makes it impossible for me to leave. "Imagine," I said, "if I have an imagination, and then work hard here, just a little every day, what might I possibly accomplish?"

Table A.4 ST and Example Trainee Translation - Perseverance

Source Text	Trainee Translation Example	Back Translation
we may follow any mood , we may look at life in fifty different ways , the only thing we must not do is to despise or deride , because of ignorance or prejudice , the influences which affect others; because the essence of all experience is that we should perceive something which we do not begin by knowing , and learn that life has a fullness and a richness in all sorts of diverse ways which we do not at first even dream of suspecting . The essayist , then , is in his particular fashion an interpreter of life , a critic of life . He does not see life as the historian , or as the philosopher , or as the poet , or as the novelist , and yet he has a touch of all these . He is not concerned with discovering a theory of it all , or fitting the various parts of it into each other . He works rather on what is called the analytic method , observing , recording , interpreting , just as things strike him , and letting his fancy play over their beauty and significance ; the end of it all being this ; that he is deeply concerned with the charm and quality of things , and gentlest light , so that at least he may make others love life a little better , and prepare them for its infinite variety and alike for its joyful and mournful surprises .	我们也许会受控于某种情绪，又或许我们会从很多不同的角度看待生活但我们唯一不能做的就是因为无知或偏见而鄙视，嘲笑他人，这种情况会影响到别人。这是由于一切经验的真谛是我们应该理解那些我们认识不到的事情，并且会从开始我们甚至会怀疑的方法中找到正确的方法来实现生活的富足。那么散文家，就是在他们翻译或评论家的世界中，得到一种独有的时尚。他不是以一位历史学家的角度看待生活，也不是站在哲学家、诗人、或是小说家的立场上，因为这些他都没有经历过。他不致力于探索以上所有这些话题的理论，也不原意为了相互融合而调整各种各样的部件。他用所谓的分析法工作，观察，记录，翻译，就像有什么在鞭策他一样，并且在美丽和壮观之上，他创立梦幻般的生活，结尾就是这样，他深深地沉浸在事物的魅力和质量上，这样做至少他可以使别人更爱生活一些，使他们为欢乐的伙食悲痛的经历做好准备。	We may be controlled by certain emotions, or we may see life from many different perspectives. But the only thing we cannot do is to despise or ridicule others because of ignorance or prejudice. This situation will affect others. This is due to the fact that the essence of all experience is that we should understand things that we don't know, and we will find the right way to achieve the enrichment of life from the beginning we may even doubt the methods. Then the essayist, in their translation or critic world, gets a unique fashion. He does not look at life from the perspective of a historian, nor does he stand in the position of a philosopher, poet, or novelist, because he has not experienced it. He is not committed to exploring the theories of all these topics, nor does he intend to adjust various components in order to integrate with each other. He works, observes, records, and translates in so-called analytics, just as what spurs him, and in beauty and spectacularness, he creates a dreamlike life. This is the end, he is deeply immersed in things. In terms of charisma and quality, at least he can make others love life more and prepare them for the sad experience of a happy meal.

Table A.5 ST and Example Trainee Translation - Essayist

Source Text	Trainee Translation Example	Back Translation
transplant surgeons work miracles . they take organs from one body and integrate them into another , granting the lucky recipient a longer , better life . sadly , every year thousands of other people are less fortunate , dying while they wait for suitable organs to be found . the terrible constraint on organ transplantation is that every life extended depends on the death of someone young enough and healthy enough to have organs worth transplanting . such donors are few . the waiting lists are long , and getting longer . freedom from this constraint is the dream of every transplant surgeon . so far attempts to make artificial organs have been disappointing : nature is hard to mimic . henec the renewed interest in trying to use organs from animals . doctors in india have just announecd that they have sucecssfully transplanted a heart from a pig into a person . pressure to increase the number of such xenotransplants seems to be growing . in europe and america , herds of pigs are being specially bred and genetically engineered for organ donation . during 1996 at least two big reports on the subject – one in europe and one in america – were published . they agreed that xenotransplants were permissible on ethical grounds . and cautiously recommended that they be allowed . america 's food and drug administration has already published draft guidelines for xenotransplantation . the ethics of xenotransplantation are relatively unworrying . people already kill pigs both for food and for sport ; killing them to save a human life seems , if anything , easier to justify . however , the scienec of xenotransplantation is much less straightforward .	器官移植 外科医生 带来 了 奇迹 。 他 们 将 器官 从 一个 身体 中 取出 并 将 它们 植入 他者 体内 ， 让 那些 有幸 得到 它们 的 人活 得 更长 ， 更好 。 令人 难过 的 是 ， 每年 都 有 数以千 计 的 人 在 等待 合适 器官 的 过程 中 死去 。 他们 就 不 那么 幸运 了 。 器官 移植术 的 可怕 的 限制 就 在于 ， 只有 那些 足够 年青 和 健康 的 死者 的 器官 才 可以 进行 移植 ， 让 生命 延续 。 这样 的 捐赠者 很少 。 而 等候 者 名单 却 很长 ， 并且 是 越来越 长 。 打破 这样 的 束缚 是 每 一个 器官移植 手术 师 的 梦想 。 迄今 为 至 制造 人工 器官 所 做 的 努力 ， 结果 却是 让 人 沮丧 的 ： 自然 难以 被 模拟 。 因此 人们 又 从 利用 动物 器官 是 寻找 到 了 希望 。 印度 的 医生 就 宣布 道 ， 他们 已经 成功 地 将 猪 的 心脏 植入 人体 。 对于 这种 与日俱 增 的 异种 器 移植 的 外界压力 也 是 越来越 大 。 在 欧美 ， 被 用来 贡献 器官 的 猪 群 被 特殊 培育 和 进行 基因 的 操纵 。 在 1996 年 ， 这方面 出 了 两个 大 报道 。 一个 在 欧洲 ， 一个 在 美国 。 他们 承认 异种 器官移植 在 伦理道德 领域 是 被 允许 的 ， 并 谨慎 地 推荐 它们 被 采纳 使用 。 美国 的 食品 药物 管理机构 已经 出版 了 异种 器官移植 草案 准则 。 这种 手术 在 伦理道德 领域 相对而言 ， 不 那么 令人担忧 了 。 人们 杀猪 来 获得 食物 ， 进行 运动 ， 那么 杀猪 去 救人 ， 就 更 解释 得通 了 。 然而 ， 异种 器官移植 科学 却 远 比 这 复杂 。	Organ transplant surgeons have brought miracles. They remove the organs from one body and implant them in the other, so that those who are fortunate enough to have them live longer and better. It is sad that thousands of people die every year while waiting for the right organs. They are not so lucky. The terrible limitation of organ transplantation is that only those organs of the deceased who are young and healthy enough can be transplanted to allow life to continue. Such donors are few. The list of waiters is long and growing. Breaking this bondage is the dream of every surgeon transplanting an organ. Until now, efforts to create artificial organs have resulted in disappointing results: Nature is difficult to simulate. Therefore, people are looking for hope from the use of animal organs. Indian doctors announced that they have successfully implanted the pig's heart in the human body. The external pressure on this growing number of foreign transplants is also growing. In Europe and the United States, pigs that are used to contribute to organs are specially cultivated and genetically manipulated. In 1996, there were two major reports in this area. One in Europe and one in the United States. They admit that xenotransplantation is allowed in the ethical field, and cautiously recommends that they be adopted for use. The US Food and Drug Administration has published draft guidelines for xenotransplantation. This kind of surgery is relatively less worrying in the ethical field. People kill pigs to get food, exercise, then kill pigs to save people, it makes even more sense. However, the science of xenotransplantation is far more complicated than this.

Table A.6 ST and Example Trainee Translation - Xenotransplantation

Appendix B

ATA Certification Programme Rubric for Grading

ATA CERTIFICATION PROGRAM
Rubric for Grading
Version 2011

Exam number: _____
Exam passage: _____

Evaluation by Dimensions

Instructions: In each column, the grader marks the box that best reflects performance in that dimension, measured against the ideal performance defined for that dimension in the "Standard" row. The grader may also insert, circle, and/or cross out words in a description to make the evaluation more specific.
Note: A passage may show uneven performance across the dimensions. For example, a candidate with excellent command of the target language but limited knowledge of the source language might show *Strong* performance for *Target mechanics* but *Minimal* performance for *Usefulness / transfer*.
See also the Explanation on the reverse.

	Usefulness / transfer	Terminology / style	Idiomatic writing	Target mechanics
STANDARD	The translated text is fully usable for the purpose specified in the Translation Instructions. The meaning and sense of the source text have been fully and appropriately transferred to the translated text.	Terminology is appropriate in context. Style and register are appropriate for the topic in the target language and for the specified audience.	Translated text reads smoothly. Wording is idiomatic and appropriate for the topic in the target language and for the specified audience.	Translated text fully follows the rules and conventions of target language mechanics (spelling, grammar, punctuation, etc.).
Strong	☐ Translated text transfers meaning in a manner fully consistent with the Translation Instructions. Translation contains few or no transfer errors, and those present have a minor effect on meaning.	☐ Translated text contains few or no inappropriate term or style/register choices. Any errors have a minor effect on meaning.	☐ Translated text is almost entirely idiomatic and appropriate in context. Any errors have a minor effect on meaning.	☐ Translated text contains few or no errors in target language mechanics.
Acceptable	☐ Translated text transfers meaning in a manner sufficiently consistent with the Translation Instructions. Translation contains occasional and/or minor transfer errors that slightly obscure or change meaning.	☐ Translated text contains occasional and/or minor inappropriate term or style/register choices. Such errors may slightly obscure meaning.	☐ Translated text contains occasional unidiomatic or inappropriate wording. Such errors may slightly obscure meaning.	☐ Translated text contains occasional errors in target language mechanics.
Deficient	☐ Translated text transfers meaning in a manner somewhat consistent with the Translation Instructions. Translation contains more than occasional transfer errors that obscure or change meaning.	☐ Translated text contains frequent inappropriate and/or incorrect terms or style/register choices. Such errors may obscure or change meaning.	☐ Translated text contains frequent and/or obvious unidiomatic or inappropriate wording. Such errors may obscure or change meaning.	☐ Translated text contains frequent and/or obvious errors in target language mechanics.
Minimal	☐ Translated text transfers meaning in a manner inconsistent with the Translation Instructions. Translation contains frequent and/or serious transfer errors that obscure or change meaning.	☐ Translated text contains excessive inappropriate and/or incorrect terms or style/register choices. Such errors obscure or change meaning.	☐ Translated text contains excessive and/or disruptive unidiomatic or inappropriate wording. Such errors obscure or change meaning.	☐ Translated text contains excessive and/or disruptive errors in target language mechanics.

ATA CERTIFICATION PROGRAM
Rubric for Grading

Explanation

Each row of the table on the reverse represents a performance level. The texts in each cell describe elements of performance at the respective level for the respective dimension. Although a candidate may perform at different levels for different dimensions, the overall usefulness of the target text can be described at general levels that correspond to the dimension indicators, where *Strong* or *Acceptable* correspond to a passing score and *Deficient* or *Minimal* represent a failing score. These overall levels are also roughly equivalent to the specific levels set forth in the Interagency Language Roundtable (ILR) Skill Level Descriptions for Translation Performance
(http://www.govtilr.org/skills/AdoptedILRTranslationGuidelines.htm).

STANDARD
The target text would require little if any editing in order to be used for the purpose specified in the Translation Instructions. (**Roughly equivalent to ILR Professional Performance Level 5**)

Strong
The target text could be published or used for professional purposes after minimal work by a bilingual editor and a target language copy editor. (**Roughly equivalent to ILR Professional Performance Level 4 or higher**)

Acceptable
A client requesting this translation could use the target text for the purpose given in the Translation Instructions after some work by a bilingual editor and/or a target language copy editor. (**Roughly equivalent to ILR Professional Performance Level 3 or 3+**)

Deficient
The target text would require extensive bilingual editing and/or target language copy editing before it could be used for the purpose given in the Translation Instructions. (**Roughly equivalent to ILR Limited Performance Level 2+**)

Minimal
This translation cannot be used for the purpose given in the Translation Instructions. It would be more economical in terms of time and money for the end user to have the text retranslated. (**Roughly equivalent to ILR Limited Performance Level 2 or lower**)

Notes:

Appendix C

High Correlation Features with Quality Scores

High correlative Features with Usefulness

Feature class	Features	r
alignment	word alignment normalized by target length	0.6
	word alignment normalized by source length	0.53
	two word alignment normalized by target length	0.49
	two word alignment normalized by source length	0.42
	TFIDF string cosine similarity with Bing translation	0.4
	target source constituency CBD	0.36
	source target discourse CBD	0.32
	bilingual embedding similarity	0.31
	target source conjunct log ratio	-0.3
	target source pronouns	-0.31
	target source basic connectives	-0.32
	target source determiner log ratio	-0.33
	target source coordination conjunction log ratio	-0.33
	target source sentence number log ratio	-0.33
	target source open clausal complement log ratio	-0.33
	target source subordinate log ratio	-0.33
	target source adjective log ratio	-0.34
	target source additives	-0.34
	target source average sentence length log ratio	-0.35
	target source punctuation log ratio	-0.35
	target source noun log ratio	-0.35
	target source adjective phrase log ratio	-0.35
	target source clausal complement log ratio	-0.36
	target source verb log ratio	-0.37

Continued on next page

Table – *Continued from previous page*

Feature class	Features	r
	target source auxiliary log ratio	-0.38
	target source clausal modifier of noun log ratio	-0.38
	target source addition connectives log ratio	-0.4
	target source adverbial log ratio	-0.4
	target source prepositional phrase log ratio	-0.41
	target source demonstratives log ratio	-0.41
	source target CBD	-0.43
	target source logical connectives log ratio	-0.43
	target source adverbial phrase log ratio	-0.46
	target source SRL others log ratio	-0.46
	target source case log ratio	-0.48
	target source marker log ratio	-0.49
	target source linkings log ratio	-0.49
	target source adjective modifier log ratio	-0.5
	target source root log ratio	-0.5
	target source noun phrase log ratio	-0.52
	target source empty word log ratio	-0.53
	target source content words log ratio	-0.53
	target source SRL A0 log ratio	-0.54
	target source SRL A1 log ratio	-0.55
	target source nominal modifier log ratio	-0.56
	target source adverbial modifier log ratio	-0.57
	target source verb phrase log ratio	-0.57
	target source nominal subject log ratio	-0.59
	source target TTR log ratio	-0.59
	source target tokens log ratio	-0.6
	source target types log ratio	-0.6
	target source punctuation dependency log ratio	-0.6
	target source conjunct log ratio	-0.6
	source target LM probability log ratio	-0.62
	target source object log ratio	-0.62
	target logical connectives	0.37
	source causal connectives	-0.3
	source linkings	-0.32
constituency	target verb phrase	0.55
	target noun phrase	0.54
	target adverbial phrase	0.5
	target prepositional phrase	0.41
	target adverbial phrase	0.41

Continued on next page

Table – *Continued from previous page*

Feature class	Features	r
	target adjective phrase	0.4
dependency	target object	0.51
	target punctuation dependency	0.5
	target nominal modifier	0.5
	target adverbial modifier	0.5
	target nominal subject	0.49
	target adjectival modifier	0.49
	target root dependency	0.48
	target phrasal verb particle	0.47
	target case	0.47
	target compound	0.45
	target maker	0.44
	target conjunct	0.42
	target clausal modifier	0.38
	target open clausal complement	0.35
	target clausal complement	0.35
	target unspecified dependency	0.35
	target copula	0.35
	target auxiliary	0.33
	source punctuation dependency	-0.33
language model	target LM probability	-0.56
POS tags	target nouns	0.56
	target verb	0.52
	target punctuation	0.51
	target adverbials	0.49
	target adjective	0.43
	target numericals	0.33
	source determiner	0.32
pseudo reference and back translation	Yandex pseudo-reference corpus level NIST score	0.48
	Yandex pseudo-reference corpus level RIBES score	0.43
	Google pseudo-reference corpus-level NIST score	0.42
	geometric mean of cosine similarities with pseudo-references	0.36
	Google pseudo-reference TFIDF weighted string cosine	0.34

Continued on next page

Table – *Continued from previous page*

Feature class	Features	r
	Yandex pseudo-reference TFIDF weighted string cosine	0.3
	Bing back translation Jaccard distance	-0.3
	Googlepseudo-reference Levenshtein distnace 1	-0.31
	Bing back translation Sorensen distance	-0.32
	Google back translation Jaccard distance	-0.32
	Yandex back translation Levenshtein 2	-0.33
	Google back translation Sorensen distance	-0.33
	Yandex back translation Levenshtein 1	-0.34
	Google back translation Levenshtein 1	-0.34
	Bing back translation Levenshtein 1	-0.34
	Yandex pseudo-reference Jaccard distance	-0.35
	Yandex pseudo-reference Sorensen distance	-0.36
	Bing pseudo-reference Jaccard	-0.37
	Bing pseudo-reference Levenshtein 1	-0.38
	Bing pseudo-reference Sorensen distance	-0.39
	Google pseudo-reference Jaccard	-0.39
	Google pseudo-reference Sorensen	-0.41
	Bing pseudo-reference Levenshtein 2	-0.41
	Yandex pseudo-reference Levenshtein 2	-0.42
	Google pseudo-reference Levenshtein 2	-0.42
	Google pseudo-reference Levenshtein 2	-0.42
	Bing pseudo-reference Levenshtein 2	-0.44
semantic roles	target SRL others	0.54
	target SRL A1	0.51
	target SRL A0	0.44
	target empty words	0.42
shallow	target type token ratio	0.62
	target tokens	0.57
	target types	0.55
	target content words	0.54
	source average sentence length	0.45

Table C.1 Contributive Features to Usefulness ($|\mathrm{r}| > 0.3$)

High correlative Features with Terminology

Feature class	Features	r
alignment	word alignment normalized by target length	0.58
	word alignment normalized by source length	0.51
	two word alignment normalized by target length	0.47
	two word alignment normalized by source length	0.41
bilingual distance	source target discourse CBD	0.39
	target source bilingual embedding similarity	0.34
	target source constituency CBD	0.3
	target source open clausal complement log ratio	-0.32
	target source unspecified dependency log ratio	-0.33
	target source auxiliary log ratio	-0.34
	target source adjective log ratio	-0.35
	target source sentence number log ratio	-0.35
	target source conjunct log ratio	-0.36
	target source determiner dependency log ratio	-0.37
	target source pronoun log ratio	-0.39
	target source noun log ratio	-0.39
	target source punctuation log ratio	-0.4
	target source determiner log ratio	-0.4
	target source adverbial phrase log ratio	-0.41
	target source clausal modifier of noun log ratio	-0.41
	target source verb log ratio	-0.42
	target source clausal complement log ratio	-0.42
	target source coordinating conjunction log ratio	-0.42
	target source basic connective log ratio	-0.43
	target source average sentence length log ratio	-0.43
	target source adverbial log ratio	-0.44
	target source addition connectives log ratio	-0.44
	target source SRL others log ratio	-0.45
	target source prepositional phrase log ratio	-0.46
	target source logical connectives log ratio	-0.47
	target source additives log ratio	-0.47
	target source demonstratives log ratio	-0.48
	target source adverbial phrase log ratio	-0.51
	target source case log ratio	-0.51
	target source linkings log ratio	-0.52
	target source root log ratio	-0.53
	target source empty word log ratio	-0.55
	target source adjectival modifier log ratio	-0.55
	source target shallow features CBD	-0.55
	target source marker log ratio	-0.56

Continued on next page

Table – *Continued from previous page*

Feature class	Features	r
	target source content log ratio	-0.57
	target source SRL A1 log ratio	-0.57
	target source SRL A0 log ratio	-0.58
	target source punctuation dependency log ratio	-0.6
	target source nominal modifier log ratio	-0.6
	target source noun phrase log ratio	-0.6
	target source conjunct log ratio	-0.62
	target source nominal subject log ratio	-0.62
	target source adverbial modifier log ratio	-0.62
	source target types log ratio	-0.64
	source target TTR log ratio	-0.64
	source target tokens log ratio	-0.64
	target source VP log ratio	-0.65
	target source object log ratio	-0.66
	source target LM probability log ratio	-0.67
cohesion and coherence	target logical connectives	0.36
	target linkings	0.35
	target additives	0.33
	target addition connectives	0.31
	target source subordinate log ratio	-0.31
	Source causal connectives	-0.35
constituency	target noun phrase	0.52
	target verb phrase	0.5
	target adverbial phrase	0.47
	target prepositional phrase	0.4
	target adverbial phrase	0.39
	target adjectival phrase	0.37
dependency	target nominal modifier	0.5
	target adjectival modifier	0.47
	target object	0.47
	target conjunct dependency	0.46
	target adverbial modifier	0.46
	target case marking	0.45
	target nominal subject	0.44
	target root dependency	0.42
	target phrasal verb particle	0.42
	target maker	0.4
	target compound	0.4
	target clausal complement	0.34

Continued on next page

Table – *Continued from previous page*

Feature class	Features	r
	target conjunct	0.34
	target coordinating conjunction	0.33
	target copula	0.32
	target open clausal complement	0.31
	target unspecified dependency	0.31
language model	target LM probability	-0.54
POS tags	target nouns	0.51
	target verbs	0.48
	target adverbials	0.46
	target punctuation marks	0.45
	target adjectives	0.41
	source adjectives	0.3
pseudo reference and back translation	Yandex pseudo-reference corpus-level NIST score	0.53
	Yandex pseudo-reference corpus-level RIBES score	0.45
	Google pseudo-reference corpus-level NIST score	0.42
	Bing pseudo-reference TFIDF weighted string cosine similarity	0.4
	geometric mean of cosine similarities with pseudo-references	0.35
	Google pseudo reference string TFIDF weighted cosine similarity	0.34
	Bing pseudo-reference TFIDF weighted cosine similarity	0.31
	Google pseudo-reference TFIDF weighted cosine similarity	0.3
	Yandex back translation Sorensen	-0.31
	Yandex pseudo-reference Levenshtein 2	-0.36
	Google pseudo-reference Levenshtein 1	-0.36
	Yandex back translation Levenshtein 1	-0.37
	Bing back translation Jaccard	-0.37
	Google back translation Jaccard	-0.38
	Bing pseudo-reference Levenshtein 1	-0.38
	Google back translation Levenshtein 1	-0.39
	Bing back translation Sorensen	-0.39
	Google back translation Sorensen	-0.39
	Bing back translation Levenshtein 1	-0.4

Continued on next page

Table – *Continued from previous page*

Feature class	Features	r
	Yandex pseudo-reference Jaccard	-0.41
	Yandex pseudo-reference Sorensen	-0.43
	Bing pseudo-reference Jaccard	-0.45
	Google pseudo-reference Levenshtein2	-0.45
	Bing back transaltion Levenshtein2	-0.46
	Google pseudo-reference Jaccard	-0.46
	Bing pseudo-reference Levenshtein 2	-0.47
	Bing pseudo-reference Sorensen	-0.47
	Google pseudo-reference Sorensen	-0.48
	Google back translation Levenshtein 2	-0.49
	Yandex back translation Levenshtein 2	-0.49
	target SRL others	0.49
semantic roles	target SRL A1	0.46
	target SRL A0	0.42
	target type token ratio	0.61
	target tokens	0.53
	target types	0.51
shallow	target content words	0.49
	target punctuation dependency	0.44
	source average sentence length	0.4
	target empty words	0.39

Table C.2 Contributive Features to Terminology ($|r| > 0.3$)

High Correlative Features with Idiomatic Writing

Feature class	Features	r
	word alignment normalized by source length	0.45
	word alignment normalized by source length	0.39
alignment	two word alignment normalized by target length	0.38
	two word alignment normalized by source length	0.34
	source target discourse CBD	0.34
	target source determiner dependency log ratio	-0.3
	target source clausal modifier of noun log ratio	-0.3
	target source unspecified dependency log ratio	-0.31
	target source adverbial phrase log ratio	-0.32
	target source SRL others log ratio	-0.33

Continued on next page

Table – *Continued from previous page*

Feature class	Features	r
	target source coordinating conjunction log ratio	-0.33
	target source average sentence length log ratio	-0.35
	target source noun log ratio	-0.35
	target source pronoun log ratio	-0.35
	target source determiner log ratio	-0.35
	target source prepositional phrase log ratio	-0.35
	target source punctuation log ratio	-0.35
	target source addition connectives log ratio	-0.36
	target source clausal complement log ratio	-0.37
	target source verb phrase log ratio	-0.37
	target source basic connective log ratio	-0.38
	target source case log ratio	-0.39
	target source adverbial log ratio	-0.4
	target source additives log ratio	-0.42
	target source logical connectives log ratio	-0.43
	source target shallow features CBD	-0.43
	target source adjectival modifier log ratio	-0.44
	target source empty word log ratio	-0.44
	target source linkings log ratio	-0.45
	target source SRL A1 log ratio	-0.46
	target source adverbial phrase log ratio	-0.46
	target source root log ratio	-0.46
	target source SRL A0 log ratio	-0.47
	target source content log ratio	-0.48
	target source marker log ratio	-0.48
	target source demonstratives log ratio	-0.49
	target source nominal modifier log ratio	-0.5
	target source nominal subject log ratio	-0.5
	target source noun phrase log ratio	-0.5
	target source punctuation dependency log ratio	-0.51
	target source conjunct dependency log ratio	-0.52
	source target types log ratio	-0.53
	source target tokens log ratio	-0.54
	source target TTR log ratio	-0.54
	target source adverbial modifier log ratio	-0.55
	target source verb phrase log ratio	-0.56
	target source object log ratio	-0.57
	source target LM probability log ratio	-0.6
	target logical connectives	0.34
cohesion and coherence		

Continued on next page

Table – *Continued from previous page*

Feature class	Features	r
	target demonstrative	0.31
	source argument type token ratio	-0.31
	source causal	-0.35
constituency	target verb phrase	0.41
	target noun phrase	0.41
	target adverbial modifier	0.41
	target prepositional phrase	0.32
	target adverbial phrase	0.32
	target nominal modifier	0.4
	target adverbial modifier	0.39
	target conjunct dependency	0.38
	target object	0.37
	target punctuation dependency	0.37
	target adjectival modifier	0.36
	target nominal subject	0.36
	target case marking	0.35
	target root	0.35
	target phrasal verb particle	0.32
	target compound	0.31
	source determiner dependency	0.31
	target marker	0.3
	Source adjective modifier	-0.3
language model	source LM perplexity	-0.3
	target LM probability	-0.45
POS tags	target adverbials	0.4
	target verbs	0.39
	target nouns	0.39
	target punctuation marks	0.38
pseudo reference and back translation	Yandex pseudo-reference NIST score	0.42
	Yandex pseudo-reference corpus-level RIBES	0.37
	Google pseudo-reference corpus level NIST score	0.31
	Bing back translation Levenshtein 2	-0.3
	Google Levenshtein2	-0.33
	Bing pseudo-reference Levenshtein 2	-0.34
	Yandex pseudo-reference Jaccard	-0.35
	Google back translation Levenshtein2	-0.36
	Bing pseudo-reference Jaccard	-0.36
	Yandex pseudo-reference Sorensen	-0.36

Continued on next page

Table – *Continued from previous page*

Feature class	Features	r
	Bing pseudo-reference Sorensen	-0.37
	Yandex back translation Levenshtein2	-0.38
	Google pseudo-reference Jaccard	-0.38
	Google pseudo-reference Sorensen	-0.39
semantic roles	target tokens	0.42
	target SRL others	0.41
	target SRL A1	0.35
	target SRL A0	0.35
shallow	target TTR	0.49
	target types	0.41
	target content words	0.39
	source average sentence length	0.33
	target empty words	0.31

Table C.3 Contributive Features to Idiomatic Writing ($|\mathrm{r}| > 0.3$)

High Correlative Features with Target Mechanics

Feature class	Features	r
alignment	word alignment normalized by target length	0.42
	word alignment normalized by source length	0.36
	two word alignment normalized by target length	0.32
	source target discourse CBD	0.34
	target source coordinating conjunction log ratio	-0.31
	target source pronoun log ratio	-0.31
	target source logical connectives log ratio	-0.31
	target source prepositional phrase log ratio	-0.31
	target source noun log ratio	-0.32
	target source verb log ratio	-0.32
	target source determiner log ratio	-0.32
	target source clausal complement log ratio	-0.33
	target source addition connectives log ratio	-0.33
	target source punctuation mark log ratio	-0.34
	target source additive log ratio	-0.35
	target source average sentence length log ratio	-0.35
	target source SRL others log ratio	-0.35
	target source adverbial log ratio	-0.35

Continued on next page

Table – *Continued from previous page*

Feature class	Features	r
	target source case log ratio	-0.36
	target source demonstratives log ratio	-0.37
	source target shallow features CBD	-0.37
	target source linkings log ratio	-0.38
	target source maker log ratio	-0.4
	target source adverbial phrase log ratio	-0.41
	target source SRL A1 log ratio	-0.41
	target source empty word log ratio	-0.42
	target source root log ratio	-0.43
	target source adjectival modifier log ratio	-0.43
	target source punctuation dependency log ratio	-0.44
	target source content log ratio	-0.46
	target source adverbial modifier log ratio	-0.46
	target source SRL A0 log ratio	-0.46
	target source conjunct dependency log ratio	-0.47
	target source nominal modifier log ratio	-0.47
	target source nominal subject log ratio	-0.47
	source target types log ratio	-0.48
	source target tokens log ratio	-0.49
	source target TTR log ratio	-0.49
	target source noun phrase log ratio	-0.49
	target source verb phrase log ratio	-0.5
	source target LM probability log ratio	-0.5
	target source object log ratio	-0.51
cohesion and coherence	source adjacent sentence overlapping	0.3
constituency	target adverbial phrase	0.35
	target verb phrase	0.34
	target noun phrase	0.33
dependency	target conjunct dependency	0.33
	target nominal modifier	0.33
	target adverbial modifier	0.33
	target adjectival modifier	0.32
	target compound	0.3
	target object	0.3
	source adjectival modifier	-0.32
language model	target LM probability	-0.34
POS tags	target noun	0.35
	target adverbvial	0.34

Continued on next page

Table – *Continued from previous page*

Feature class	Features	r
pseudo reference and back translation	Google back translation Levenshtein 2	-0.32
	Yandex back translation Levenshtein 2	-0.33
semantic roles	target SRL others	0.33
shallow	target type token ratio	0.41
	target content	0.36
	target tokens	0.34
	target types	0.32

Table C.4 Contributive Features to Target Mechanics ($|\mathrm{r}| > 0.3$)

Hight Correlative Features with Adequacy

Feature class	Features	r
alignment	word alignment normalized by target length	0.62
	word alignment normalized by source length	0.54
	two word alignment normalized by target length	0.5
	two word alignment normalized by source length	0.43
bilingual distance	source target discourse CBD	0.36
	target source constituency CBD	0.35
	Source target bilingual embedding similarity	0.33
	target source unspecified dependency log ratio	-0.31
	target source determiner dependency log ratio	-0.33
	target source conjunct dependency log ratio	-0.33
	target source subordinate log ratio	-0.34
	target source open clausal complement	-0.34
	target source sentence number log ratio	-0.35
	target source pronoun log ratio	-0.35
	target source adjective log ratio	-0.35
	target source determiner log ratio	-0.37
	target source basic connective log ratio	-0.37
	target source clausal complement log ratio	-0.38
	target source auxiliary log ratio	-0.38
	target source noun log ratio	-0.38
	target source punctuation log ratio	-0.38
	target source adjective phrase log ratio	-0.39

Continued on next page

Table – *Continued from previous page*

Feature class	Features	r
	target source average sentence length log ratio	-0.39
	target source clausal complement log ratio	-0.4
	target source verb log ratio	-0.4
	target source clausal modifier of noun log ratio	-0.4
	target source additive log ratio	-0.4
	target source addition connectives log ratio	-0.43
	target source adverbial log ratio	-0.43
	target source prepositional phrase log ratio	-0.44
	target source demonstratives log ratio	-0.45
	target source logical connectives log ratio	-0.46
	target source SRL others log ratio	-0.47
	source target shallow features CBD	-0.49
	target source adverbial phrase log ratio	-0.49
	target source case log ratio	-0.51
	target source linkings log ratio	-0.52
	target source root log ratio	-0.53
	target source adjectival modifier log ratio	-0.53
	target source marker log ratio	-0.53
	target source empty word log ratio	-0.55
	target source content log ratio	-0.56
	target source noun phrase log ratio	-0.57
	target source SRL A1 log ratio	-0.57
	target source SRL A0 log ratio	-0.57
	target source nominal modifier log ratio	-0.6
	target source adverbial modifier log ratio	-0.61
	target source nominal subject log ratio	-0.62
	target source punctuation dependency log ratio	-0.62
	target source verb phrase log ratio	-0.62
	source target TTR log ratio	-0.63
	target source conjunct dependency log ratio	-0.63
	source target tokens log ratio	-0.63
	source target types log ratio	-0.63
	target source object log ratio	-0.65
	source target LM probability log ratio	-0.66
cohesion and coherence	target logical	0.38
	target linkings	0.32
	source linkings	-0.31
	source causal connective log ratio	-0.33
	target noun phrase	0.55

Continued on next page

Table – *Continued from previous page*

Feature class	Features	r
constituency	target verb phrase	0.55
	target adverbial phrase	0.51
	target prepositional phrase	0.42
	target adverbial phrase	0.42
	target adjective phrase	0.4
	target nominal modifier	0.52
	target object	0.52
	target adverbial modifier	0.5
	target punctuation dependency	0.5
	target adjectival modifier	0.5
	target nominal subject	0.49
	target case marking	0.48
	target root	0.48
	target phrasal verb particle	0.47
	target conjunct dependency	0.45
	target compound	0.45
	target marker	0.44
	target clausal modifier of noun	0.38
	target copula	0.35
	target open clausal comlement	0.35
	target unspecified dependency log ratio	0.34
	target clausal complement	0.34
	source determiner dependency	0.33
	target auxiliary	0.31
	source punctuation dependency	-0.31
language model	target LM probability	-0.57
POS tags	target nouns	0.56
	target verbs	0.52
	target punctuation marks	0.51
	target adverbial	0.5
	target adjective	0.44
	target numeral	0.32
	Yandex pseudo-reference NIST score	0.52
	Yandex pseudo-reference corpus level RIBES score	0.45
	Google pseudo-reference corpus level NIST score	0.43
	Bing pseudo-reference TFIDF weighted cosine similarity	0.41

Continued on next page

Table – *Continued from previous page*

Feature class	Features	r
	geometric mean of cosine similarities with pseudo-references	0.37
	Google pseudo-reference TFIDF weighted string cosine similarity	0.35
	Yandex pseudo-reference string cosine similarity	0.3
	Google pseudo-reference Levenshtein 1	-0.34
	Bing back translation Jaccard	-0.34
	Yandex back translation Levenshtein 2	-0.35
	Google back translation Jaccard	-0.35
	Bing back translation Sorensen	-0.35
	Yandex back translation Levenshtein 1	-0.36
	Google back translation Sorensen	-0.37
	Bing pseudo-reference Levenshtein 1	-0.37
	Google pseudo-reference Levenshtein 1	-0.37
	Yandex pseudo-reference Jaccard	-0.38
	Bing back translation Levenshtein 1	-0.4
	Yandex pseudo-reference Sorensen	-0.4
	Bing pseudo-reference Jaccard	-0.42
	Bing pseudo-reference Sorensen	-0.43
	Google pseudo-reference Jaccard	-0.43
	Bing pseudo-reference Levenshtein 2	-0.44
	Google pseudo-reference Levenshtein 2	-0.45
	Google pseudo-reference Sorensen	-0.45
	Yandex back translation Levenshtein 2	-0.46
	Google back translation Levenshtein 2	-0.46
	Bing pseudo-reference Levenshtein 2	-0.47
semantic roles	target SRL others	0.54
	target SRL A1	0.51
	target SRL A0	0.45
shallow	target type token ratio	0.64
	target tokens	0.57
	target types	0.55
	target content	0.54
	source average sentence length	0.45
	target empty words	0.42

Table C.5 Contributive Features to Adequacy ($|\mathrm{r}| > 0.3$)

High Correlative Features for Fluency

Feature class	Features	r
alignment	word alignment normalized by target length	0.45
	word alignment normalized by source length	0.39
	two word alignment normalized by target length	0.37
	two word alignment normalized by source length	0.33
bilingual distance	source target discourse CBD	0.35
	target source determiner dependency log ratio	-0.3
	target source clausal modifier of noun log ratio	-0.3
	target source adverbial phrase log ratio	-0.32
	target source coordinating conjunction	-0.34
	target source pronoun log ratio	-0.35
	Target SRL other log ratio	-0.35
	target source noun log ratio	-0.35
	target source prepositional phrase log ratio	-0.35
	target source determiner log ratio	-0.35
	target source punctuation mark log ratio	-0.36
	target source average sentence length log ratio	-0.36
	target source clausal complement log ratio	-0.36
	target source verb log ratio	-0.37
	target source basic connective log ratio	-0.37
	target source case log ratio	-0.39
	target source adverbial log ratio	-0.4
	source target shallow features CBD	-0.42
	target source empty words log ratio	-0.45
	target source adjectival modifier log ratio	-0.45
	target source SRL A1 log ratio	-0.46
	target source adverbial phrase log ratio	-0.46
	target source demonstratives log ratio	-0.46
	target source root log ratio	-0.46
	target source marker log ratio	-0.47
	target source SRL A0 log ratio	-0.49
	target source content log ratio	-0.49
	target source punctuation dependency log ratio	-0.5
	target source nominal modifier log ratio	-0.5
	target source nominal subject log ratio	-0.51
	target source noun phrase log ratio	-0.52
	target source conjunct dependency log ratio	-0.52
	source target types log ratio	-0.53

Continued on next page

Table – *Continued from previous page*

Feature class	Features	r
	source target tokens log ratio	-0.54
	target source adverbial modifier log ratio	-0.54
	source target TTR log ratio	-0.54
	target source verb phrase log ratio	-0.56
	target source object log ratio	-0.57
	source target LM probability log ratio	-0.58
cohesion and coherence	target demonstrative	0.3
	source causal connective	-0.33
	target source addition	-0.36
	target source logical log ratio	-0.4
	target source additive log ratio	-0.41
	target source linkings log ratio	-0.44
constituency	target adverbial phrase	0.4
	target verb phrase	0.4
	target noun phrase	0.4
	target prepositional phrase	0.3
dependency	target nominal modifier	0.39
	target adverbial modifier	0.38
	target conjunct dependency	0.38
	target object	0.36
	target adjectival modifier	0.36
	target case marking	0.34
	target nominal subject	0.34
	target logical connectives	0.34
	target root	0.34
	target compound	0.32
	target phrasal verb particle	0.31
	target marker	0.3
	source adjectival modifier	-0.32
language model	target LM probability	-0.42
POS tags	target adverbials	0.39
	target nouns	0.39
	target verb	0.37
	target punctuation marks	0.36
	target punctuation marks	0.35
pseudo reference and back translation	Yandex pseudo-reference corpus level NIST score	0.39
	Yandex pseudo-reference RIBES score	0.34
	Yandex pseudo-reference Jaccard	-0.31

Continued on next page

Table – *Continued from previous page*

Feature class	Features	r
	Google pseudo-reference Levenshtein 2	-0.31
	Bing pseudo-reference Levenshtein 2	-0.32
	Yandex pseudo-reference Sorensen	-0.33
	Bing pseudo-reference Jaccard	-0.34
	Google pseudo-reference Jaccard	-0.35
	Bing pseudo-reference Sorensen	-0.35
	Google back translation Levenshtein 2	-0.36
	Google pseudo-reference Sorensen	-0.36
	Yandex back translation Levenshtein 2	-0.37
semantic roles	target SRL others	0.4
	target SRL A0	0.34
	target SRL A0	0.32
shallow	target TTR	0.47
	target tokens	0.4
	target content	0.39
	target types	0.39
	source average sentence length	0.33

Table C.6 Contributive Features to Fluency ($|\mathrm{r}| > 0.3$)

High Correlative Features with Total Score

Feature class	Features	r
alignment	word alignment normalized by source length	0.58
	word alignment normalized by source length	0.51
	two word alignment normalized by target length	0.47
	two word alignment normalized by source length	0.41
	source target discourse CBD	0.37
	target source constituency CBD	0.33
	Target source bilingual embedding similarity	0.32
	target source unspecified dependency log ratio	-0.31
	target source open clausal complement log ratio	-0.33
	target source subordinate log ratio	-0.33
	target source conjunct dependency log ratio	-0.33
	target source sentence number log ratio	-0.33
	target source determiner log ratio	-0.33
	target source adjective log ratio	-0.34

Continued on next page

Table – *Continued from previous page*

Feature class	Features	r
bilingual distance	target source auxiliary log ratio	-0.35
	target source pronoun log ratio	-0.36
	target source adverbial phrase log ratio	-0.38
	target source coordinating conjunction log ratio	-0.38
	target source determiner log ratio	-0.38
	target source clausal modifier of noun log ratio	-0.38
	target source basic connective log ratio	-0.38
	target source noun log ratio	-0.38
	target source punctuation marks log ratio	-0.39
	target source average sentence length log ratio	-0.4
	target source clausal complement log ratio	-0.4
	target source verb log ratio	-0.4
	target source additive log ratio	-0.42
	target source addition connectives log ratio	-0.42
	target source prepositional phrase log ratio	-0.43
	target source adverbial log ratio	-0.43
	target source SRL others log ratio	-0.45
	target source logical connectives log ratio	-0.46
	target source demonstratives log ratio	-0.47
	source target shallow features CBD	-0.48
	target source case log ratio	-0.49
	target source adverbial phrase log ratio	-0.5
	target source linkings log ratio	-0.51
	target source root log ratio	-0.53
	target source adjectival modifier log ratio	-0.53
	target source marker log ratio	-0.53
	target source empty word log ratio	-0.54
	target source SRL A1 log ratio	-0.55
	target source content log ratio	-0.56
	target source SRL A0 log ratio	-0.56
	target source noun phrase log ratio	-0.57
	target source nominal modifier log ratio	-0.59
	target source punctuation dependency log ratio	-0.6
	target source nominal subject log ratio	-0.6
	target source adverbial modifier log ratio	-0.61
	target source conjunct dependency log ratio	-0.62
	source target TTR log ratio	-0.62
	source target types log ratio	-0.62

Continued on next page

Table – *Continued from previous page*

Feature class	**Features**	**r**
	target source verb phrase log ratio	-0.62
	source target tokens log ratio	-0.62
	target source object log ratio	-0.65
	source target LM probability log ratio	-0.66
cohesion and coherence	target logical	0.38
	target linkings	0.33
	source causal connective	-0.34
constituency	target noun phrase	0.52
	target verb phrase	0.52
	target adverbial phrase	0.49
	target prepositional phrase	0.39
	target adverbial phrase	0.39
	target adjective phrase	0.36
dependency	target nominal modifier	0.49
	target object	0.48
	target adverbial modifier	0.48
	target adjectival modifier	0.47
	target punctuation dependency	0.47
	target nominal subject	0.45
	target case marking	0.45
	target root	0.45
	target conjunct dependency	0.44
	target phrasal verb particle	0.43
	target compound	0.42
	target marker	0.41
	target clausal modifier of noun	0.35
	target unspecified dependency	0.33
	target open clausal comlement	0.32
	source determiner dependency	0.32
	target copula	0.31
language model	target LM probability	-0.54
POS tags	target noun	0.52
	target verb	0.49
	target adverbial	0.48
	target punctuation marks	0.47
	target adjective	0.4
	Yandex pseudo-reference corpus-level NIST score	0.49

Continued on next page

Table – *Continued from previous page*

Feature class	Features	r
pseudo reference and back translation	Yandex pseudo-reference corpus-level RIBES score	0.43
	Google pseudo-reference corpus-level NIST score	0.4
	Bing pseudo-reference TFIDF weighted string cosine	0.38
	geometric mean of cosine similarities with pseudo-references	0.33
	Google pseudo-reference TFIDF weighted string cosine	0.32
	Bing back translation Jaccard	-0.31
	Google pseudo-reference Levenshtein 1	-0.32
	Yandex pseudo reference Levenshtein 2	-0.32
	Google back translation Jaccard	-0.33
	Bing back translation Sorensen	-0.33
	Google back translation Sorensen	-0.34
	Bing pseudo-reference Levenshtein 1	-0.34
	Google back translation Levenshtein 1	-0.35
	Yandex back translation Levenshtein 1	-0.35
	Bing back translation Levenshtein 1	-0.36
	Yandex pseudo-reference Jaccard	-0.37
	Yandex pseudo-reference Sorensen	-0.39
	Bing pseudo-reference Jaccard	-0.4
	Bing back translation Levenshtein 2	-0.41
	Google pseudo-reference Levenshtein 2	-0.42
	Google pseudo-reference Jaccard	-0.42
	Bing pseudo-reference Sorensen	-0.42
	Bing pseudo-reference Levenshtein 2	-0.43
	Google pseudo-reference Sorensen	-0.44
	Google back translation Levenshtein 2	-0.44
	Yandex back translation Levenshtein 2	-0.44
semantic roles	target SRL others	0.51
	target SRL A1	0.47
	target SRL A0	0.42
shallow	target type token ratio	0.6
	target tokens	0.53
	target types	0.52

Continued on next page

Table – *Continued from previous page*

Feature class	**Features**	$\mathbf{r}$
	target content	0.51
	source average sentence length	0.42
	target empty words	0.39

Table C.7 Contributive Features to Total ($|\mathrm{r}| > 0.3$)

Appendix D

Features Unique to Document-Level Translations

Feature Class	Features
Cohesion and Coherence	averaged source bag of words LSA cosine averaged source word embedding cosine averaged source word embedding pearson averaged embedding correlation distance averaged target embedding cosine averaged target emedding pearson averaged target embedding correlation distance target source LSA cosine log ratio target source embedding cosine log ratio target source embedding pearson log ratio target source embedding correlation distance log ratio source adjacent sentence overlapping target adjacent sentence overlapping target source content empty CBD source sentence number target sentence number target source sentence number log ratio

Table D.1 Features unique to document-level translations

Appendix E

Lexicon of English and Chinese Connectives

	English Connectives	**Chinese Connectives**
	English Connectives	Chinese Connectives
Basic	for, and, nor, but, or, yet so	因 为,由 于,对 于,和,及,并,以及,与,就,而且,但是,然后,而,且,及其,亦不,也不,也不是,也没有,但是,而 是,然 而,仅 仅,只,或,或者,还 是,但 是,然 而,所 以,因此,仍,尚,还,且,犹 自,可 是,却,仍旧,或 是,亦 或,抑 或,要 么,要 不然,只是,不过
subordinator	after, although, as, because, before, if, once, since,that, though, till, unless, until,whenever, wherever, whereas, whereupon, while	后,一 旦,为 止,之 前,于 是,以后,以 致 于,倘,倘 及,倘 若,假如,假 若,先 于,兹 因,凡 是,则,即使,即 便,却,因,因 为,如,如 同,如果,如 若,尽 管,届 时,必 要 时,无论,无 论 何 处,无 论 何 时,既 是,既然,是 否,每 当,每 逢,然 后,然而,由于,皆因,直到,直至,纵 使,纵然,而,至,若,若,若 是,虽,虽 则,虽是,虽然,虽说,要是,身为,鉴于,随着

Continued on next page

Table – *Continued from previous page*

	English Connectives	**Chinese Connectives**
addition	and, also, besides, further, furthermore, too, moreover, addition,then, another, indeed, likewise	和,与,和, 敢 情,与,乃,也,也是,亦,以及,何况,其外,其实,再,再则,再 者,再 说,况 且,又,及,另,另一,另 一 个,另 外,同 样,实在,并,并,并 且,愈 加,接 着,方才,更 有 甚 者,此 外,然 而,继,继而,而且,而后,诚然,还,还有,进一步,遂,那时,除了,除开,竟
linkings	nonetheless, therefore, although, furthermore, whereas, nevertheless, whatever,for, however,besides, henceforth, then, yet, if, while, so, but, until,because,alternatively, meanwhile,when, and, since, notwithstanding, whenever,moreover, as, with, consequently, after	尽 管 如 此,一 旦,不 管,不 管,不论,不 过,且,为 止,乃,于 是,于是 乎,今 后,仍,仍 旧,仍 然,从此,从 而,以 及,以 后,任 何,但,但是,倒,倘,倘 及,倘 或,倘 然,倘若,假如,假若,其外,再则,再者,再说,况,况 且,凡,则,加 之,加 以,即使,却,却是,及,另一,另外,可是,同时,后,和,因,因,因为,因此,因此,因而,好歹,如同,如果,如若,对于,尚且,就,尽 管,并,并 且,当,当 年,当时,必 要 时,怎 么 着,总 要,或,所以,接着,故,故此,故而,方才,无论何事,无论如何,既,既是,既然,是以,是否,是故,更有甚者,果若,此外,每 当,然 后,然 而,犹,由 于,皆因,直到,直至,结果,继,继而,而,而且,自 此,至,苟,若,若 是,若 然,虽则,虽是,虽然,虽说,要是,设若,身为,还是,还有,这样,那时,鉴于,除了,随着,饶
order	in conclusion, next, first, firstly, second, secondly, finally,to begin with, above all,before, after,then	首先, 第二,接下来,其后,在此之后,第一,其一, 第二,最后,总之,之前,在之前,以后,之后,然后

Continued on next page

Table – *Continued from previous page*

	English Connectives	**Chinese Connectives**
reason-purpose	therefore, that is why, for this reason, for that reason, hence,because, so,since, as, because of,on account of, so that, consequently	因此,所以,故,故此,因而,是以,乃,是故,于是乎,为此,这就是为什么,正因为如此,这也是为什么,由于这一原因,出于这个原因,为此原因,鉴于上述原因,于是,故此,因,由于,兹因,以便,足以,好使,以致于
opposition	but, however, nevertheless, otherwise, on the other hand, on the contrary,yet,still, maybe, perhaps,instead,except for, in spite of, despite, nonetheless, apart from, unlike,whereas	但,除外,不一定,不一样,不同,不然,不过,不顾,与之相反,之外,仍,仍然,以外,但,但是,依然,兴许,却,反之,反而,反过来说,另一方面,只是,可是,可能,尚,尚且,尽管,恐怕,或,或者,然而,相反,纵使,而,而是,莫不是,虽然,要不,要不然,说不定,还是,除,除了,除去
demonstrative	this, that, these, those	这,本,此,这种,这,那,那样,那个,彼,这些个,该等,这些,那些,那些个

Continued on next page

Table – *Continued from previous page*

	English Connectives	**Chinese Connectives**
additive	after all, again, all in all, also, alternatively, and, anyhow, as a final point,as well, at least, besides, but, by contrast, by the way, contrasted with, correspondingly, except that, finally,first, for example, for instance, fortunately, further,furthermore, however, in actual fact, in addition, in contrast, in fact, in other words, in sum, incidentally,instead, it follows,moreover,next, notwithstanding that,on one hand, on the contrary, on the one hand, on the other hand, or, otherwise, rather, secondly, similarly,summarizing,summing up, that is, thereupon, to conclude,to return to, to sum up, to summarize, to take an example,to these ends,to this end,too,well at any rate,whereas,yet,	总归,一方面,下一,不然,不管,不过,与,与之不同的是,与此不同,与此对照,与此相反,且,为例,为此,举一个例子,举个例子,举例来说,乃,之外,也就是说,也就是说,事实上,事实上,于是,亦,亦即,亦或,以及,但,但是,何况,例如,值得庆幸的是,其实,其实不然,其次,再者,再说,最后,最后一点是,况且,到目前为此,即使,却,反之,反而,反观,另一方面,另外,可是,同样,同样地,和,实际上,尚未,就是说,就此,尽管,并,并且,幸好,幸运的是,庆幸的是,归根到底,归纳,当然,形成鲜明对比的是,总之,总体上说,总结,总结,总而言之,意外地,或,或不,或是,或者,所幸的是,换句话说,换言之,接下来,无论如何,末了,概括,概述,此外,比如,比方说,毕竟,然而,由此可见,相反,相应,相应地,相比之下,相比而言,第一,第二,类似于,终究,综上所述,而,而不是,而且,至今,至少,至此,要不,要不然,譬如,话又说回来,话说回来,起码,迄今,迄今为此,近似,还,还有,进一步,遂,除外,除开,除非,随即,顺便一提,顺便提一下,顺便提一句,顺便说一下,顺带一提,首先

Continued on next page

Table – *Continued from previous page*

	English Connectives	**Chinese Connectives**
causal	although, arise, arises, arising, arose, because, cause, caused, causes, causing,condition,conditions, consequence, consequences, consequent, consequently, due to, enable, enabled, enables, enabling, even then, follow that,follow the, follow this,followed that, followed the, followed this, following that, follows the, follows this,hence, made, make, makes, making, nevertheless, nonetheless, only if, provided that,result, results, since, so, therefore, though,thus, unless, whenever	为条件,之后,之后,于是,产生,仅当,从而,令,以使,以便,但是,使,使得,使成,倘若,出于,出现,前提,即使是这样,即使这样,即便如此,发生,只有,只有在,只有当,只要,后果,因为,因此,因此可以说,因而,导致,尽管,尽管如此,引出,引发,引起,必要时,惟有,所以,按照,故,既然,是故,每当,然而,由于,由此,结果,继,致使,虽则,虽然,话又说回来,话说回来,诱发,遵循,除非,随即,随后的

Continued on next page

Table – *Continued from previous page*

	English Connectives	**Chinese Connectives**
logical	actually, admittedly, after all, all in all, also, alternatively, although,and conversely, anyhow, anyway,arise from, arise out of,arises from, arises out of,arising from, arising out of, arose from, arose out of,as a final point, as a result, as well, at least, at this point,because, besides, but, by contrast, cause, caused,causes, causing, conditional upon, consequence, consequences, consequently, contrasted with, correspondingly, despite the fact that,due to,enable, enabled, enables, except that,finally, follow that, follow the,follow this, followed that, followed the,followed this,following that, follows the, follows this, for, fortunately, further, furthermore,hence	实际,一句话,一方面,万一,下,下一,下一个,下次,不少于,不然,不管,不管怎样,不过,与之相反,与此同时,与此对照,与此相反,为了,为例,为实现这一目的,为条件,为此,为此目的,主要目的是,举个例子来说,举例来说,乃,之后,也,也不,也就是,也就是说,也没有,事实上,于是乎,亏得,亦不,亦或,产生,产生于,令,以使,以便,任何情况下,但,但对于,但是,使,使得,使成,依据,倒反,倘若,假使,假如,其他,其外,其实,其实不然,其次,其目标是,再不,再次,再者,再说,最后,最少,出乎,出于,即使,即使是这样,即使这样,即便如此,却,原因是,及,反之,反倒,反而,反过来,反过来说,另,另一方面,另外,只要,可是,同样,同样地,后果,否则,回到,因,因为,因此,因而,在这一点上,好在,如果,实际上,尚要,就是说,尽管,尽管如此,幸好,幸而,幸运,幸运的是,庆幸的是,引发,引起,归根到底,归根结蒂,当前,当然,形成对照的是,形成强烈对比,形成鲜明对比,得以

Continued on next page

Table – *Continued from previous page*

	English Connectives	**Chinese Connectives**
	however, if, in actual fact,in any case, in any event, in case, in conclusion,in contrast, in fact, in order that, in other words, in short, in sum,incidentally,instead, it followed that, it follows, it follows that, likewise, moreover, nevertheless,next,nor, nonetheless,notwithstanding that, on condition that, on one hand,on the condition that, on the contrary, on the one hand, on the other hand, once again, or, otherwise, provided that, purpose of which, pursuant to, rather,secondly, similarly, since, so, summarizing,summing up, that is, that is to say,then, therefore, thereupon, though, thus, to conclude, to return to, to sum up,to summarize,to take an example,to that end, to these ends, to this end, to those ends, unless, well at any rate, whereas,while	总之,总体上说,总体而言,总归,总的来说,总结,总而言之,恰恰相反,惟有,愈加,意外地,或,或是,或者,所以,所以,所幸的是,抑或,按照,换句话说,换言之,故此,故而,无可否认,无论如何,既是,既然,易言之,是以,是因为,是故,有别于,条件下,条件为,条件是,根据,概括,概要,概述,此刻,此外,此时,此时此刻,毕竟,然则,然后,然而,现在,由于,由引可见,目前,相反,相反地,相对于,相对而言,相应地,相比之下,相比较之下,相比较而言,立刻,第二,简言之,类似的,纵使,纵然,结果是,继,继而,而,而是,能够,至少,致使,虽然,虽然,虽然,要不,要不然,要么,诚然,话又说回来,话说回来,诱发,还,还好,还是,进一步,进行对比,造成,遵循,限便如此,除了,除外,除开,除非,随之,随即,随后,随后的,顺便一提的是,顺便指出,顺便提一句,顺便说一下,顺便说一句,顺带一提,须

Table E.1 Bilingual Lexicon of English and Chinese Connectives