Document-Level Machine Translation Quality Estimation

A Thesis submitted to the University of Sheffield for the degree of Doctor of Philosophy in the Faculty of Engineering

by

Carolina Scarton

Department of Computer Science

University of Sheffield

September 2016

Acknowledgements

I would like to acknowledge the EXPERT project (EXPlointing Empirical appRoaches to Translation - Marie Curie FP7 ITN n. 317471) for the financial support, the opportunity to networking with the brightest minds in the Machine Translation area and for supporting and encouraging internships.

Special thanks must go to my supervisor, Lucia Specia, for her guidance and friendship throughout the last three years. Lucia is always kind and positive, which makes the difference during the "dark periods" of the PhD life. She also gave me unconditional support from correcting my English writing mistakes to discussing crucial research ideas, without which this PhD would not be possible. It has been an amazing opportunity to work with her and her team in Sheffield.

Special thanks to Rob Gaizauskas and Eleni Vasilaki, that were members of my PhD panel committee, for their extremely helpful advice at the early stages of my research. I am also thankful to my PhD examiners, Marco Turchi and Mark Stevenson, for making the viva process a fruitful chat that resulted in the improvement of this thesis.

Many thanks to Constantin Orasan for the hard work done in organising the EXPERT project, and to all senior and early stage researches involved in this project, mainly to Josef van Genabith, from Saarland University in Germany, and Manuel Herraz and Alex Helle, from Pangeanic in Spain, for supervising me during my secondments. Special thanks must go to Liling, Rohit, José Manuel, Marcos, Mihaela, Peggy, Stefania, Karin, Andrea, Katrin, Ashraf, Katja and Jörg from Saarland University for the collaboration and friendship during my time in Saarbrücken, and to Varvara, Hannah, Hoang and Carla, EXPERT early stage and experienced researches that I had the opportunity to interact with in Sheffield.

I am also very thankful to my Master's supervisor Sandra Maria Aluísio and to Magali Sanches Duran, from the University of São Paulo, for all the support and collaboration along all these years. I would also like to thank all teachers from the Professora Alva Fabri Miranda and Colégio Técnico Industrial schools and lectures from the University of São Paulo that played an important role in my education.

Many thanks to my colleagues in the Natural Language Processing group for making the work environment very pleasant. Especially to Fred, Kashif, Daniel, Gustavo, Roland, Karin, David, Andreas and Ahmet for all the conversations about research and non-related topics, all the coffee breaks and all after-work drinks.

Many thanks to my friends Kim, Wilker, Marialaura, Miguel, Victor, Teo, Tiago, Claudia, Jorge, Angela, Raluca, Marta, Umberto, Eletta, Carol and Fer for making part of my unforgettable time outside Brazil. I would also like to thank my jiu-jitsu team at The 5 Rings Grappling Academy (The Forge), especially my coaches John, Gregg and Paulinho. Special thanks must also go to Lara, João, Iara, Raquel, Camila, Diego, Shimizu, Bigão, Tiago, Boulos, Vilma, Luiz, Bete, Silvana, Célio and Eduardo, my Brazilian friends that, apart from the distance, were always there for me.

Many many thanks to my future husband, Brett, for all the patience, support and love. I would also like to thank Natalie, Michael, Cyril and Gisselle (Brett's family) for accepting me as part of their family, which helps a lot when I am homesick. Special thanks must go to my parents Luis and Lourdes and my brother Daniel for all the love, for believing in me and for supporting me in pursuing my dreams even though it kept us apart. Many thanks also to Rita, Sidnei, Sofia, Pedro, Vó Dalva, Vó Maria, Natalia, Carlos, Inês, Paulo and Lucas (my family) for all the support and love during these three years. Finally, very special thanks to my beloved granddad Antonio for everything that I have learnt from him and for his love during the time he was here with us.

Abstract

Assessing Machine Translation (MT) quality at document level is a challenge as metrics need to account for many linguistic phenomena on different levels. Large units of text encompass different linguistic phenomena and, as a consequence, a machine translated document can have different problems. It is hard for humans to evaluate documents regarding document-wide phenomena (e.g. coherence) as they get easily distracted by problems at other levels (e.g. grammar). Although standard automatic evaluation metrics (e.g. BLEU) are often used for this purpose, they focus on *n*-grams matches and often disregard document-wide information. Therefore, although such metrics are useful to compare different MT systems, they may not reflect nuances of quality in individual documents.

Machine translated documents can also be evaluated according to the task they will be used for. Methods based on measuring the distance between machine translations and post-edited machine translations are widely used for task-based purposes. Another task-based method is to use reading comprehension questions about the machine translated document, as a proxy of the document quality. Quality Estimation (QE) is an evaluation approach that attempts to predict MT outputs quality, using trained Machine Learning (ML) models. This method is robust because it can consider any type of quality assessment for building the QE models. Thus far, for document-level QE, BLEU-style metrics were used as quality labels, leading to unreliable predictions, as document information is neglected. Challenges of document-level QE encompass the choice of adequate labels for the task, the use of appropriate features for the task and the study of appropriate ML models.

In this thesis we focus on feature engineering, the design of quality labels and the use of ML methods for document-level QE. Our new features can be classified as document-wide (use shallow document information), discourse-aware (use information about discourse structures) and consensus-based (use other machine translations as pseudo-references). New labels are proposed in order to overcome the lack of reliable labels for document-level QE. Two different approaches are proposed: one aimed at MT for assimilation with a low requirement, and another aimed at MT for dissemination with a high quality requirement. The assimilation labels use reading comprehension questions as a proxy of document quality.

The dissemination approach uses a two-stage post-editing method to derive the quality labels. Different ML techniques are also explored for the document-level QE task, including the appropriate use of regression or classification and the study of kernel combination to deal with features of different nature (e.g. handcrafted features versus consensus features). We show that, in general, QE models predicting our new labels and using our discourse-aware features are more successful than models predicting automatic evaluation metrics. Regarding ML techniques, no conclusions could be drawn, given that different models performed similarly throughout the different experiments.

Table of contents

List of figures

List of tables

List of Acronyms

ACT Accuracy of Connective Translation

AS Automatic Summarization

BLEU BiLingual Evaluation Understudy

BP Brevity Penalty

CBOW Continuous Bag-of-Words

CCG Combinatory Categorical Grammar

CE Confidence Estimation

CREG Corpus of Reading Comprehension Exercises in German

CSLM Continuous Space Language Model

DLPT Defence Language Proficiency Test

EDU Elementary Discourse Unit

GP Gaussian Process

GTM General Text Matcher

HTER Human-targeted Translation Edit Rate

LC Lexical Cohesion

LDA Latent Dirichlet Allocation

LM Language Model

LSA Latent Semantic Analysis

LSI Latent Semantic Indexing

LSTM Long Short-Term Memory

MAE Mean Absolute Error

METEOR Metric for Evaluation of Translation with Explicit ORdering

ML Machine Learning

MT Machine Translation

NIST National Institute of Standards and Technology

NLP Natural Language Processing

NP Noun Phrase

PCFG Probabilistic Context-Free Grammar

PDTB Penn Discourse Treebank

PE Post-Editing

PEA Post-Editing Action

POS part-of-speech

QE Quality Estimation

RA Readability Assessment

RBMT Rule-based MT

ROUGE Recall-Oriented Understudy for Gisting Evaluation

RST Rhetorical Structure Theory

RTM Referential Translation Machines

SCFG Synchronous Context-Free Grammar

SMT Statistical Machine Translation

SRL Semantic Role Labelling

SVD Singular Vector Decomposition

SVM Support Vector Machines

SVR Support Vector Regression

TER Translation Error Rate

TF-IDF Term Frequency - Inverse Document frequency

TOEFL Test of English as a Foreign Language

TOEIC Test Of English for International Communication

TM Translation Memory

TS Text Simplification

WMT Conference on Machine Translation

WSD Word Sense Disambiguation

Chapter 1

Introduction

A major challenge in Natural Language Processing (NLP) is **to find ways to evaluate language output tasks** such as Machine Translation (MT), Automatic Summarization (AS) and Text Simplification (TS). Although the nature of these tasks is different, they are related in the sense that a "target" text is produced given an input "source" text. Evaluation metrics for these tasks should be able to measure quality with respect to different aspects (e.g. fluency and adequacy) and should be scalable across different systems and datasets. Human evaluation is the most desirable approach, but it presents several drawbacks. Firstly, human evaluation is not immune to biases where humans would give scores based on their perception of automatic systems, for example. Secondly, this kind of evaluation is time-consuming, expensive and not available for on-demand cases (such as applications targeting directly the end-user). Finally, for some cases, humans can get confused and bored during the evaluation process, which makes it unreliable. Therefore, a significant amount of work has targeted measuring quality of language output tasks without direct human intervention.

In the MT area, the focus of this thesis, machine translated texts are mainly used in two scenarios: dissemination and assimilation (Nirenburg, 1993). The **dissemination** scenario concerns machine translated texts with the purpose of publication. An example is a news agency that may want to make their online English content available for readers of other languages. For that, the quality requirements are high and the final version of the translation is often edited and revised by humans.

MT can also be used for information **assimilation**. For example, scientists around the world may wish to know the latest findings by the scientific community in Brazil about the Zica virus by machine translating Portuguese articles. In this case, the quality of the machine translated documents does not need to be perfect, as long as the text is understandable and the meaning of the source is preserved. More recently, with the broad availability and use of the

internet and social media, there is a third scenario, where the aim is **communication**. In this case, users apply MT technologies with the purposes of exchanging information, chatting, dating, ordering goods from foreign countries, among others. As with the assimilation scenario, the quality requirements are low: only the most important information needs to be understandable.

Therefore, a reliable evaluation framework should ideally take into account the purpose of the translation, the targeted audience and the type of the documents, among other aspects. Human evaluation would be the most desirable, although it has several problems, as mentioned before, and, therefore, automatic evaluation metrics have been proposed to overcome the issues with human evaluation. The BiLingual Evaluation Understudy (BLEU) (Papineni et al., 2002), the Translation Error Rate (TER) (Snover et al., 2006) and the Metric for Evaluation of Translation with Explicit ORdering (METEOR) (Banerjee and Lavie, 2005) are widely used automatic evaluation metrics for MT. These metrics compare the outputs of MT systems with human reference translations. BLEU is a precision-oriented corpus-based metric that compares n-grams (typically $n = 1..4$) from reference texts against n-grams in the MT output, measuring how close the output of a system is to one or more references. TER measures the minimum number of edits required to transform the MT output into the closest reference texts at sentence level. METEOR scores MT outputs by aligning their words with words in a reference. The alignments can be done by exact, stem, synonym and paraphrase matchings at sentence level.

One limitation of these automatic metrics is that if the MT output is considerably different from the references, it does not really mean that it is a bad output. Another problem is that human effort is still needed to produce the references. Also, the matching or alignment methods are simplistic. For example, all errors are considered equally important (e.g. a wrong comma is as important a wrong main verb in a sentence) and it is generally not possible (or cumbersome) to customise the metrics for different purposes. Finally, and more importantly, these metrics cannot be applied in cases where the output of the system is to be used directly by end-users. For example, a user reading the output of Google Translate[1] for a given news text cannot count on a reference for that translated text.

Alternatively, the output of an MT system can be evaluated based on its applications. For example, post-editing can be used as a proxy to evaluate the effort of correcting a machine translation (Plitt and Masselot, 2010; Blain et al., 2011; O'Brien, 2011; Specia et al., 2011; Koponen et al., 2012). Therefore, human interaction is needed to correct the machine translation in order to achieve the same meaning as the source and ensure the fluency and

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

style of the translation. Human-targeted Translation Edit Rate (HTER) (Snover et al., 2006) is a metric that indirectly evaluates the effort needed to post-edit a sentence. It is calculated by applying TER between the machine translation and the post-edited machine translation (i.e., it replace the human reference with the post-edited version). The post-editing approaches for evaluation are normally linked to the use of MT dissemination purposes, where the quality to be delivered needs to be high.

Another task-based approach is the evaluation of machine translated documents using reading comprehension questions. The hypothesis behind this approach is that if humans are able to answer questions about a machine translated document, the document is considered as good, if not, the document is considered as bad (Tomita et al., 1993; Fuji, 1999; Fuji et al., 2001; Jones et al., 2005b,a; Berka, Černý, and Bojar, 2011). This approach is often used to evaluate MT for the assimilation purpose, where an understandable version of the document is enough.

A more recent form of MT evaluation is referred to as Quality Estimation (QE). QE approaches aim to predict the quality of MT systems without using references. Features (that may or may not be related to the MT system that produced the translations) are extracted from source and target texts (Blatz et al., 2004; Specia et al., 2009a; Bojar et al., 2013, 2014, 2015, 2016b). The only requirement is data points with quality scores (e.g. HTER or even BLEU-style metrics) to train supervised Machine Learning (ML) models in order to predict the quality of unseen data. The advantage of these approaches is that they only need human intervention in order to produce enough data points to build the ML model for QE and, therefore, unseen data does not need to be manually annotated. QE systems predict scores that reflect how good a translation is for a given scenario and, therefore, can be considered a task-oriented evaluation approach. For example, HTER scores are widely used as quality labels for QE systems at sentence level, providing a measure of post-editing effort. A user of a QE system predicting HTER could decide whether to post-edit or translate sentences from scratch based on the scores predicted for each sentence.

So far, QE has been largely explored for word and sentence levels, with little work on document-level QE. Sentence-level QE (Specia et al., 2009a,b; Specia and Farzindar, 2010; Felice and Specia, 2012; Shah, Cohn, and Specia, 2013; Shah et al., 2015a) is the most explored of the three levels, with direct applications on the translation industry workflow. Word-level QE (Blatz et al., 2004; Ueffing and Ney, 2005; Luong, 2014) aims to give a quality score for each word in the translation. Its application include spotting errors, where systems could inform users about the quality of individual words. Both sentence and word levels have been continously explored at the Conference on Machine Translation (WMT),

through QE shared tasks. Document-level QE (Soricut and Echihabi, 2010; Scarton and Specia, 2014a; Scarton, 2015; Scarton et al., 2015) aims to predict a single score for entire documents, which has proven to be a hard task even for humans (Scarton et al., 2015).

In this thesis we focus on document-level QE because it has been less explored than other levels, it is useful for a number of applications and it has several challenges commonly found in other NLP tasks (e.g. AS).

Document-level QE can be used to evaluate machine translations of entire documents that need to be used "as is", without post-editing (assimilation case). A scenario would be quality assessment of machine translated user generated content in e-services that deal with buyers from all around the world. For example, an English speaking user searching for a hotel in Greece might be interested in the reviews of other people that stayed in this hotel. However, the majority of the reviews are in Greek and, therefore, machine translating the reviews is the only option for the user who does not speak Greek. In this scenario, the user might be interested in the overall quality of the machine translated review that he/she is reading, in order to decide if he/she can trust it or not. An evaluation of quality per sentence is not useful, because the user is interested in the full document (review). In a similar case, e-services providers could also be interested in using MT to translate reviews from a language into another when they are starting to offer a product in a new country. Quality assessment of entire reviews is needed in order to select which reviews will be understood by the users and thus can be published.

Another application is to estimate the cost of post-editing machine translated documents (dissemination case). It is common for translation companies to hire freelance translators to perform translation tasks. Frequently, due to cost issues or disclosure agreements, the translators only receive a set of shuffled sentence for post-editing. Then, after receiving the post-editions, an in-house translator needs to revise the entire document, and correct document-wide problems. Such a complex scenario makes it difficult to clearly estimate the cost of a translation service.Therefore, a document-level cost estimation that takes into account the costs of the work performed by the freelancer (at sentence level) and the revision work would be useful for the translation industry.

We address the following challenges faced by QE for MT at document level:

Features Designing and identifying the best features for building the QE models is a challenge in QE. For document-level, the state-of-the-art features are based on pseudo-references (Soricut and Echihabi, 2010; Soricut, Bach, and Wang, 2012; Soricut and Narsale, 2012; Shah, Cohn, and Specia, 2013). **Pseudo-references** are translations produced by MT

systems other than the system we want to predict the quality of. They are used as "artificial" references to evaluate the output of the MT system of interest via traditional automatic evaluation metrics (e.g. BLEU). However, such features usually cannot be extracted in real-world scenarios since they make certain assumptions about the MT systems used to produce the pseudo-references (such as quality) that are often unavailable. Therefore, the design and evaluation of new features is needed. Another challenge in feature engineering for document-level QE is the use of linguistic information that goes beyond words or sentences. **Discourse** is a type of linguistic information that often manifests itself document-wide. Since the state-of-the-art MT systems translate documents at sentence-level, disregarding discourse information, it is expected that the outputs of these systems will contain discourse problems. Because of that, recently there have been initiatives to include discourse information in MT (Marcu, Carlson, and Watanabe, 2000; Carpuat, 2009; LeNagard and Koehn, 2010; Zhengxian, Yu, and Guodong, 2010; Meyer and Popescu-Belis, 2012; Ture, Oard, and Resnik, 2012; Ben et al., 2013; Hardmeier, 2014), MT evaluation (Giménez and Màrquez, 2009; Giménez et al., 2010; Meyer et al., 2012; Wong and Kit, 2012; Guzmán et al., 2014) and also in QE at sentence level (Rubino et al., 2013). However, thus far there are no contribution on document-level QE that explores document-wide or discourse-aware features and effective ways of combining such features along with more shallow information.

Quality labels Another challenge in document-level QE is devising a quality score to predict. Previous research has used automatic evaluation metrics as quality labels for document-level QE (Soricut and Echihabi, 2010; Soricut, Bach, and Wang, 2012; Soricut and Narsale, 2012). However, our hypothesis is that traditional metrics, developed to evaluate outputs of different MT systems of the same source text, do not capture differences among machine translations of different documents, because they only capture generic errors that, although are useful for system-level evaluation, are not distinctive in terms of individual document quality (n-grams matching and word alignments, for example). This is even more problematic if the documents are translated by the same or similar MT system(s). This leads to low variation between the document quality scores and, therefore, all document scores are close to the average quality of the dataset. Another problem is that automatic evaluation metrics do not account for document-wide and discourse-aware problems, they are limited to superficial information about n-grams. Finally, automatic evaluation metrics are not targeted at a purpose and cannot be directly interpreted as an absolute quality indicator by the end-user (e.g. what does a BLEU score of 0.6 mean?). Therefore, new quality labels need to be investigated in order to further develop the area of document-level QE.

ML models Investigating and developing ML models for QE is highly dependent on the task and on the quality labels provided. For instance, labels following a continuous distribution are more suitable to be used with regression models, whilst labels that follow a discrete distribution are more likely to be approached with classification models. Moreover, features of different nature (e.g. word embeddings versus handcrafted) may need to be treated differently inside the ML model. Therefore, different ML approaches need to be investigated in order to develop reliable studies in QE.

1.1 Aims and Objectives

The aims and objectives of this thesis are:

1. Investigate novel shallow and deep information sources and ways of combining these sources for the task of document-level QE. More specifically:
   (a) Sentence-level information and ways of aggregating it for document-wide approaches.
   (b) Latent variable models (e.g. Latent Semantic Analysis (LSA) (Landauer, Foltz, and Laham, 1998)) for modelling (shallow) discourse-aware features;
   (c) Linguistic theories (e.g. Rhetorical Structure Theory (RST) (Mann and Thompson, 1987)) and tools available for modelling (deep) discourse-aware features;
   (d) Consensus of MT systems as features for document-level QE;
   (e) Word embeddings as features for document-level QE;
2. Devise reliable quality labels for the task of document-level QE. More specifically:
   (a) Devise and acquire human-targeted task-oriented labels for assimilation purposes using a method based on reading comprehension tests for data collection and linear combination approaches for devising the labels;
   (b) Devise and acquire human-targeted task-oriented labels for dissemination purposes using a method based on two-stage post-editing for data collection and linear combination approaches for devising the labels;
   (c) Understand how the documents are distinguished by the label and whether or not they capture discourse-aware phenomena.

3. Investigate appropriate ML models and techniques for QE at document-level. More specifically:

   (a) Explore different approaches for modelling document-level QE: multiclass and ordinal classification and non-linear and bayesian regression;

   (b) Investigate kernel combination approaches as a way of dealing of the peculiarities of features of different nature.

1.2 Contributions

This thesis introduces the following main contributions:

- A new method for document-level evaluation for dissemination purposes called two-stage post-editing. This method consists in post-editing sentences in two steps: firstly sentences are post-edited in isolation, without the influence of document context. The idea is to solve all sentence-level issues. After this stage, the post-edited sentences are put into document context and the same post-editor is asked to perform any remaining changes.[2] These two stages aim to isolate document-aware problems and provide a resource for a more reliable document evaluation;

- New approaches for devising quality labels for document-level QE from the two-stage post-editing method, aiming to penalise documents with more document-aware issues;

- Two new methods for devising quality labels for document-level QE from reading comprehension tests (dissemination scenario), one evaluating documents by using open questions and another with multiple choice questions. In the first case, the document scores follows a continuous distribution, given that the open questions are marked following a continuous scale. Moreover, the question marks are linearly combined and weighted by the number of questions per document. In the second case, the multiple choice questions produces a discrete distribution, where there is an ordinal relation between the scores;

- Design and analysis of new feature sets for document-level QE. The new feature sets focus on different information types: document-aware (sentence-level information

[2] It is important to have the same translator to perform both steps in order to avoid changes related to translation style.

aggregated at document-level), discourse-aware (document-wide discourse information) and consensus (pseudo-reference-based). Features are analysed in terms of their correlation with quality scores and their effectiveness in building document-level QE models;

- Annotated corpora:
  - FAPESP corpus with documents machine translated by three different MT systems from English into Brazilian Portuguese (2,823 documents);
  - WMT corpus organised by documents for English into German, Spanish and French (and vice-versa) with the purpose of document-level QE (474 documents for each language pair);
  - WMT paragraph-level corpus for the paragraph-level QE shared task organised in WMT15 (1,215 paragraphs);
  - A two-stage post-editing corpus that is a sample of WMT corpus in which the two-stage post-editing method was applied (208 documents);
  - CREG-mt-eval corpus, based on the CREG corpus (Ott, Ziai, and Meurers, 2012), this corpus was translated by three different MT systems with an extra version with mixed sentences (one from each MT systems) and human translations for some documents as references (215 documents);
  - MCtest-mt-eval, based on the MCtest corpus (Richardson, Burges, and Renshaw, 2013), with backward translation from English into German and back into English (660 documents).
- An extension of the QUEST++ framework for document-level QE. The document-level module of QUEST++ was developed based on the existing sentence-level structures. A document is considered a group of sentences and thus, 69 sentence-level features were adapted for document level. Moreover, nine new discourse features were also added into the framework;
- The organisation of WMT15 paragraph-level QE shared task and WMT16 document-level QE shared task. In the WMT15, the task consisted in predicting METEOR scores of paragraphs. The data used were taken from the WMT13 translation shared task. In WMT16, the task was predicting the quality of entire documents. Quality scores were devised from the two-stage post-editing method. The documents were selected from WMT08, 09, 10, 11, 12 and 13 translation shared tasks.

1.3 Published Material

Some parts of this thesis were published in the following:

- Carolina Scarton, Daniel Beck, Kashif Shah, Karin Sim Smith and Lucia Specia (2016): Word embeddings and discourse information for Quality Estimation. In the Proceedings of the First Conference on Statistical Machine Translation, Berlin, Germany, pp. 831-837.

- Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor and Marcos Zampieri (2016): Findings of the 2016 Conference on Machine Translation. In the Proceedings of the First Conference on Statistical Machine Translation, Berlin, Germany, pp. 131-198.

- Carolina Scarton and Lucia Specia (2016): A Reading Comprehension Corpus for Machine Translation Evaluation. In the Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, pp. 3652-3658.

- Carolina Scarton and Lucia Specia (2015): A quantitative analysis of discourse phenomena in machine translation. Discours - Revue de linguistique, psycholinguistique et informatique, number 16.

- Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia and Marco Turchi (2015): Findings of the 2015 Workshop on Statistical Machine Translation. In the Proceedings of the Tenth Workshop on Statistical Machine Translation, Lisbon, Portugal, pp. 1-46.

- Carolina Scarton, Liling Tan and Lucia Specia (2015): USHEF and USAAR-USHEF participation in the WMT15 QE shared task. In the Proceedings of the Tenth Workshop on Statistical Machine Translation, Lisbon, Portugal, pp. 336-341.

- Lucia Specia, Gustavo Henrique Paetzold and Carolina Scarton (2015): Multi-level Translation Quality Prediction with QuEst++. In the Proceedings of ACL-IJCNLP 2015 System Demonstrations, Beijing, China, pp. 110-120.

- Carolina Scarton (2015): Discourse and Document-level Information for Evaluating Language Output Tasks. In the Proceedings of NAACL-HLT 2015 Student Research Workshop (SRW), Denver, CO, pp. 118-125.

- Carolina Scarton, Marcos Zampieri, Mihaela Vela, Josef van Genabith and Lucia Specia (2015): Searching for Context: a Study on Document-Level Labels for Translation Quality Estimation. In the Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT 2015), Antalya, Turkey, pp. 121-128.

- Carolina Scarton and Lucia Specia (2014): Document-level translation quality estimation: exploring discourse and pseudo-references. In the Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT 2014), Dubrovnik, Croatia, pp. 101-108.

- Carolina Scarton and Lucia Specia (2014): Exploring Consensus in Machine Translation for Quality Estimation. In the Proceedings of the Ninth Workshop on Statistical Machine Translation (WMT 2014) - in conjunction with ACL 2014, Baltimore-MD, pp. 342-347.

Two tutorials (products of this thesis) were also given and one tutorial is being prepared:

- QUEST++ tutorial preparation and presentation at Alicante University, Alicante, Spain - 24/01/2016;[3]

- QUEST++ tutorial preparation and presentation at the International Conference on the Computational Processing of the Portuguese Language (PROPOR 2016), Tomar, Portugal - 13/07/2016;[4]

- QUEST++ tutorial preparation to be presented at the 26th International Conference on Computational Linguistics (COLING 2016), Osaka, Japan - 11/12/2016.[5]

1.4 Structure of the Thesis

In Chapter 2 we present a literature review of MT evaluation. Different kinds of evaluation approaches are discussed including manual evaluation, reference-based evaluation and human-targeted evaluation. This chapter also contains the detailed description of the QE task,

[3] http://staffwww.dcs.shef.ac.uk/people/C.Scarton/resources.html
[4] http://propor2016.di.fc.ul.pt/?page_id=705
[5] http://coling2016.anlp.jp/tutorials/T4/

including a discussion about all levels of prediction (word, sentence and document). Features for QE are also discussed, focusing on work that used linguistic information for QE at word and sentence levels.

Chapter 3 contains the literature review about discourse research for MT. We use a taxonomy of discourse (Stede, 2011) and categorise the research papers following it. A discussion of how each level is related to this thesis is presented in the end of each section.

Chapter 4 presents the document-level features used in this thesis. We separate the features in three classes: document-aware, discourse-aware and consensus features. In the end of this chapter we present a preliminary analysis of the correlation of discourse features with HTER.

In Chapter 5 we show our first experiments with document-level QE. BLEU, TER, METEOR and HTER are used as quality labels for document-level prediction. Three different datasets are used: FAPESP (English-Brazilian Portuguese), WMT (English into German, Spanish and French and vice-versa) and LIG (French into English). Finally, we discuss the results and evaluate the usefulness of automatic evaluation metrics as quality labels for document-level QE.

Chapter 6 presents our new labels, proposed in order to better evaluate documents for the task of document-level QE. Two approaches are introduced: one based on reading comprehension questions and another on a two-stage post-editing method. We then analyse the results achieved by the new labels by comparing them with automatic evaluation metrics results and discuss their effectiveness.

Finally, in Chapter 7 we summarise the thesis and provide a discussion of future directions for document-level QE research.

Chapter 2

Machine Translation Evaluation

Assessing the quality of documents is a challenge for many NLP tasks, starting from the question of defining **quality**. MT quality assessment is a subjective task that depends on various factors, including the purpose of the translation: what the text will be used for and by whom. Traditional MT evaluation uses generic metrics of error/correctness, focusing on the fact that machine translation sentences are likely to contain errors (Koehn, 2010). This kind of evaluation is useful for system comparisons and ranking of systems, but does not provide meaningful information on quality for the end-user.

As discussed in Chapter 1, automatically translated texts are mainly used in two scenarios: dissemination and assimilation (Nirenburg, 1993). The **dissemination** scenario concerns machine translated texts with the purpose of publication and, therefore, the quality requirements are high. For **assimilation**, the quality requirements are less strict: just the main information needs to be understandable.

In this chapter we describe the main evaluation approaches employed in MT. Section 2.1 presents human evaluation of machine translated texts. The settings evaluate sentences according to fluency and adequacy scores or rank MT systems.

Section 2.2 presents the use of automatic metrics for MT evaluation. Such automatic metrics, usually performed at sentence or corpus level, are reference-based: quality scores are acquired by evaluating similarities between the machine translated texts and human references. BLEU, METEOR and TER are the widely used metrics of this kind and are usually also employed for tuning MT systems.

In Section 2.3, task-based approaches for MT evaluation are discussed. Such approaches are useful if the purpose of the MT output goes beyond system evaluation and system tuning. Post-editing, the task of changing the MT output in order to achieve fluent and adequate translations, is used with the purposes of reducing costs of the translation process. For

example, a metric that evaluates the cost of post-editing a sentence (or a document) is useful to inform users about the indirect quality of the machine translation. Another task-based approach uses reading comprehension tests about the machine translated documents. In this case, the purpose is to evaluate whether or not the machine translation is comprehensible. A quality score can be devised by counting the correct answers that a human scored for each document. Finally, eye tracking techniques have also been explored to assess MT quality for task-based approaches.

In Section 2.4 we present QE: a kind of evaluation that focuses on predicting the quality of unseen data, by using models trained with data points labelled for quality. The general framework for QE is also presented, showing the modules for feature extraction from source and target texts and machine learning. QE is widely explored at sentence and word levels, with a considerable amount of work done in terms of feature engineering and ML research. Here, we focus on document-level QE, which is the main topic of this thesis. The creation of fully automated solutions that provide machine translated content directly to the end user is an example of the utility of document-level QE. Previous work on QE for sentence and word levels using linguistic features is also discussed in order to provide some background for our study on discourse information in Chapter 4. Finally, this section also includes the description of QUEST++, a tool with feature extraction and ML modules for QE, and our contribution for the document-level QE module.

In the remainder of this thesis we use the terms "target" or "target text" as a synonym of "MT output". The terms "source" or "source text" will refer to the input text given to the MT system.

2.1 Manual Evaluation

Manually evaluating NLP tasks is probably the most intuitive way of assessment that one can think of. In MT, evaluating by hand whether or not a translation is good is an approach for quality evaluation. Such evaluations can be performed by bilingual evaluators (who can judge if the target preserves the source information) and monolingual evaluators (that need a reference translation in order to assess the machine translation). Traditionally, humans evaluate MT output at sentence level, although they may benefit from the full document in order to assess discourse problems (Koehn, 2010).

Human translation evaluation is a difficult task, since different translators can have different views on the same translation and, therefore, have different translation preferences. Manually evaluating machine translations is also problematic: humans can be biased by their

own preferences and, in consequence, the judgements variation between annotators is usually high. Moreover, different subjects can have different expectations about what the purpose of the translation is and, even if the guidelines are clear whether the purpose is dissemination or assimilation, these concepts are vague and humans will probably disagree in their evaluation. Finally, human evaluation is costly, time-consuming and, therefore, cannot be performed in real-time scenarios (e.g. *gisting*).

A kind of manual evaluation that has been largely employed in MT area assesses *fluency* and *adequacy* (Koehn, 2010). Humans judgments of fluency check if a translation is fluent in the target language, i. e. if it uses the correct grammar and idiomatic choices, disregarding its meaning. On the other hand, human judgements of adequacy check if a translation preserves the same meaning as the source (or reference(s)) (Snover et al., 2009). Both fluency and adequacy are often evaluated at sentence level in a 1 to 5 scale (1 being worst and 5 the best) (Koehn and Monz, 2006; Callison-Burch et al., 2007).

Although the concepts of fluency and adequacy can be considered easy to interpret, assessing them is not trivial. Humans have different perspectives of such concepts and, therefore, the scores from different evaluators can show high variation. Especially for adequacy, humans are capable of filling in the missing information without noticing problems with the machine translation (Koehn, 2010). Moreover, Fomicheva and Specia (2016) argue that, for human evaluation scenarios using reference translations only, annotators are heavily biased by the structure of the reference(s), giving bad scores for machine translations that highly differ from the reference(s), even though the meaning is preserved.

Another widely used human evaluation technique is to rank machine translation systems. For each source sentence, a fixed number of machine translations from different MT systems (normally five) are shown to the annotators and they are asked to rank the machine translations in a fixed scale, usually from 1 (best) to 5 (worst). Therefore, the ranking task encompasses the comparison of the machine translated sentences among themselves and against the source sentence (in some cases, a human reference of the sentence under evaluation can also be available). Ties are allowed, given that different MT systems can output the same or very similar machine translations. Ideally, the same set of sentences is evaluated by more than one annotator so that agreement scores can be computed. Koehn (2010) claims that this kind of evaluation is more consistent than the fluency/adequacy scores. This type of evaluation is the official evaluation of the MT shared tasks of WMT since 2008 (Callison-Burch et al., 2008, 2009, 2010, 2011, 2012; Bojar et al., 2013, 2014, 2015, 2016b).

Both fluency/adequacy scores and ranking cannot be directly extended to document level. Judging documents directly is unfeasible to humans, since documents encompass different

problems on different levels. Averaging sentence-level scores in order to have a document level scores is probably the most intuitive way. However, as we discuss later in Chapter 6, averaged sentence-level scores are not reliable in expressing document quality.

2.2 Automatic Evaluation Metrics

In order to overcome the shortcomings of manual evaluation (such as time and costs), **automatic evaluation metrics** have been proposed over the years. Such metrics automatically assess machine translations by comparing them to one or more human translations (called reference translations). Although this approach still requires human interaction, it is much less time-consuming and expensive, since it requires texts to be translated by humans only once and the same documents can be used to evaluate different MT systems.

Similarly to other NLP tasks, MT can also be evaluated by using **precision**, **recall** and ***f-measure***. Precision is calculated as the ratio between the number of correct word matches (between target and reference) divided by the length of the target. Recall is the ratio between the number of correct word matches and the length of the reference. *F-measure* is the harmonic mean between the two metrics. Such metrics do not consider word order and they can be easily misleading. Short machine translations are more likely to show higher precision while recall can be maximised by having more repeated correct words on the machine translation than on the reference.

Therefore, more sophisticated metrics have been proposed in order to overcome the issues with precision and recall. BLEU, TER and METEOR are examples of widely used metrics for MT evaluation, although several other metrics have been proposed over the years. WMT annually organises a shared task on MT evaluation since 2008 (Callison-Burch et al., 2008, 2009, 2010, 2011, 2012; Macháček and Bojar, 2013, 2014; Stanojević et al., 2015; Bojar et al., 2016a).

Such automatic evaluation metrics are at sentence level. By aggregating sentence-level scores, a corpus level evaluation is achieved, the corpus being the test corpus (that can be composed of random disconnected sentences). For MT (mainly Statistical Machine Translation (SMT)) the context of the sentences is not important, given that the translation is performed sentence-by-sentence, disregarding context. Consequently, the traditional MT evaluation used to evaluate MT systems usually also deals with corpus made of random sentences. Therefore, automatic evaluation metrics are designed for **system evaluation and comparison** mainly and not for **absolute quality assessment of translations**. Finally, it is

common to refer to this kind of evaluation performed by such metrics as **segment-level** and **system-level** evaluation.

Although recent advances include the use of discourse information (Joty et al., 2014) and sophisticated Long Short-Term Memory (LSTM)-based approaches (Gupta, Orasan, and van Genabith, 2015) for MT evaluation, there is no approach that addresses document-level assessment apart from aggregating sentence-level scores. The reasons for this are: (i) the majority of MT systems (mainly traditional SMT systems) perform translation sentence-by-sentence and, therefore, sentence-level evaluation is still important in this scenario; (ii) the aggregation of sentence-level scores lead to a **system-level** evaluation, mainly because the entire corpus can be composed by random sentences, and (iii) the evaluation of such metrics is done against manual evaluation (ranking) which is also done at sentence level.

Additionally, the traditional assessment procedure to evaluate automatic metrics performance relies on human rankings (Stanojević et al., 2015). Humans are asked to rank sentences translated by different MT systems. These rankings are then used to evaluate the automatic metrics, by correlation scores (such as Pearson r). As such, metrics are designed and often optimised to compare different systems.

In this section we present BLEU, TER and METEOR in detail, since these metrics are very popular, perform reasonably well and are used in the experiments presented in this thesis.

BLEU

BiLingual Evaluation Understudy (BLEU) is the most widely used metric for MT evaluation. This metric, proposed by Papineni et al. (2002), is a precision-oriented metric that also takes recall into account (penalising target sentences which are shorter than the references). It was designed to be used with more than one reference, although it is also possible to be used for cases where only one reference is available. Despite the fact that sentence-level approximations are also usually performed for evaluation at segment level, BLEU is a corpus-based metric that uses information from the entire corpus.

The precision in BLEU is a **modified n-gram precision** where, for each target n-gram, it is computed the maximum number of times that this n-gram appear in any of the reference translation sentences. The total count of each candidate n-gram is then clipped by its maximum counts in the reference corpus. After that, the clipped values are summed and

divided by the total number of candidate *n*-grams in the entire corpus. Equation 2.1 shows how modified *n*-gram precision is calculated for the entire corpus.

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{\text{n-gram} \in C} Count_{clip}(\text{n-gram})}{\sum_{C' \in \{Candidates\}} \sum_{\text{n-gram'} \in C'} Count(\text{n-gram'})} \tag{2.1}$$

Although Equation 2.1 already penalises long candidate sentences, short sentences could wrongly maximise the precision. The solution proposed in BLEU was to multiply the Equation 2.1 by a factor called *Brevity Penalty (BP)* that is expected to be 1.0 if the candidate's sentence length is higher than all reference sentence length. BP is calculated for the entire corpus (Equation 2.2). First the closest values for matching reference and candidate sentences are summed for all sentences (r). Then, r/c is used in a decaying exponential equation (c being the total length of the candidate corpus).

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{1-r/c} & \text{if } c \leq r \end{cases} \tag{2.2}$$

Equation 2.3 shows the final BLEU equation (where N is the n-gram order and $w_n = 1/N$). Traditional BLEU uses $N = 4$ (4-gram).

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{2.3}$$

Since BLEU is a precision-oriented metric, values range from 0 to 100, being 0 the worst and 100 the best.

METEOR

The Metric for Evaluation of Translation with Explicit ORdering (METEOR) was designed in order to overcome BLEU's weaknesses, such as the lack of an explicit recall component, the use of *n*-grams with high order, the lack of explicit word matches and not providing a reliable sentence-level score (Banerjee and Lavie, 2005). METEOR computes explicit unigram alignments between words from target and reference translations. For cases with more than one reference, the reference that leads to the best final METEOR score is chosen.

In order to compare target and reference sentences, METEOR establishes alignments among the words. Each word in the target sentence must be aligned to zero or only one word in the reference. Such alignments are acquired in two stages. In the first stage, all possible

word-level alignments are retrieved by an external module. Such module can consider the following steps: "exact match" (where the word in the target should be exactly the same as in the reference), "stem match" (where stems are matched instead of words), "synonym match" (where words in the target can be matched to their synonyms in the reference) and "paraphrase match" (where entire phrases are matched, if they are defined as paraphrases in an external resource provided) (Denkowski and Lavie, 2014). Each step is applied in isolation and the order that they are applied matters. For example, if "exact match" is applied first and the word "dog" in the target was aligned to an exact match in the reference, this word will not be aligned again in the "stem match", even though other alignments are possible.

In the second stage, the word alignments are selected according to the position of the words in both target and reference. In fact, what is expected from this stage is to penalise target sentences that show word ordering far from the expected in the reference. METEOR, then, accounts for the number of alignment *crosses* between the words and select the set of alignments that shows less *crosses*.

With the selected word alignments, the METEOR score is generated using a harmonic mean (F_{mean} between precision and recall scores) (Equation 2.4). Precision (P) is the number of unigram alignments between target and reference divided by the number of words in the target and recall (R) is the number of alignments between target and reference divided by the number of words in the reference.

$$F_{mean} = \frac{P \cdot R}{\alpha \cdot P + (1-\alpha) \cdot R} \tag{2.4}$$

Finally, a penalty factor is added in order to take into account n-gram alignments. The number of alignments between chunks of words (n-grams) is divided by the number of word alignments, as shown in Equation 2.5.

$$Penalty = \gamma \cdot \left(\frac{ch}{m}\right)^{\beta} \tag{2.5}$$

The final METEOR score is defined by Equation 2.6. Parameters γ, β and α can be optimised to maximise correlation with human ranks.[1]

$$\text{METEOR} = (1 - Penalty) \cdot F_{mean} \tag{2.6}$$

[1] In the recent versions of METEOR other parameters can be optimised in order to calculate weighted precision and recall and take into account differences between content and function words.

METEOR is an *f-measure*-oriented metric and its values range from 0 to 100, where higher values mean better performance.

TER

Translation Error Rate (TER) (Snover et al., 2006) is a metric for MT evaluation that calculates the minimum number of edits needed to transform the target into one of the references. It was designed to be a more intuitive score when compared to BLEU and METEOR.

The **minimum number of edits** is computed only between the target and the reference that requires less edits to be achieved. Such modifications (edits) can be insertion, deletion, substitution or shifts (when words change position into the sentence). Punctuations are also considered as words and the metric is often case-sensitive. The different type of edits are treated with equal costs. The minimum number of edits is then divided by the average length of the references (all references are included in the average) (Equation 2.7).

$$\text{TER} = \frac{\text{minimum \# of edits}}{\text{average \# of reference words}} \tag{2.7}$$

Since it is not computationally possible to calculate the optimal edit-distance with shifts (such problem is *NP-Complete*), TER uses approximations calculated in two steps. Firstly, dynamic programming is used to compute the number of insertions, deletions and substitutions. The set of shifts that most reduces the number of insertions, deletions and substitutions is obtained by using greedy search. Secondly, a minimum edit distance is used to calculate the remaining edit distance (the optimal is achieved by dynamic programming).

TER is an error-based metric, with values ranging from 0 to 100, where higher values lead to worse results.

2.3 Task-based Evaluation

As mentioned before, manual and automatic evaluation metrics are mainly designed to assess system performance in comparison to other systems. However, the purpose of MT can be different: one can be interested in whether or not a machine translation output "is good enough" for a given task. MT can be used for improving translation productivity and *gisting*, for example. In both cases, a task-oriented evaluation is more informative than scores related to number of errors in sentences according to human references.

In this section, we present two different task-oriented approaches for MT evaluation. Cost or effort of post-editing is useful when MT is being used in the translation workflow and the translation quality is required to be high (therefore, it is a scenario of dissemination) (Section 2.3.1). The use of Reading Comprehension tests aim to evaluate whether a machine translated text can be understandable even though it presents errors (assimilation) (Section 2.3.2).

2.3.1 Post-editing Effort

Post-editing is the task of checking and, when necessary, correcting machine translations. Such corrections are performed by humans[2] and the hypothesis is that correcting a machine translated text is faster than translating it from scratch. Globalisation and the need for information in different languages as fast as possible gave a higher importance for translation solutions that goes beyond the traditional human translation workflow. Consequently, there is a need to make the translation process faster and more accurate (Dillinger, 2014). Post-editing of machine translations is one approach associated to the use of MT for dissemination.

However, as expected, it is not always the case that correcting a machine translation is quicker than translating it from scratch. Some sentences have such low quality that the task of reading it, trying to understand it and correcting it is more time-consuming than translating it from scratch. Therefore, estimating post-editing effort to support the translator work in automatic ways (e.g. informing the translator whether or not it is worth post-editing a sentence) can be an informative metric in the translation process.

According to Krings (2001), post-editing effort has three dimensions: temporal, cognitive and technical. The temporal dimension is the most straightforward of the three. It is the direct measurement of the time spent by the post-editor to transform the MT output into a good quality post-edited version. Although cognitive aspects are directly related to temporal effort, they cannot be fully captured directly. Cognitive aspects encompass linguistic phenomena and style patterns; and their measurements can only be done by using indirect means of effort assessment (e.g. keystrokes pauses). For example, a simple change in a verb tense require much less cognitive effort than resolving an anaphora. Finally, the technical dimension involves the practical transformations performed in order to achieve the post-edited version. Such transformations can be insertion, deletion, shift or a combination of all of them. It is worth noting that the technical dimension focuses on the different operations without

[2]Although initiatives to automate post-editing already exists (Bojar et al., 2015), here we only refer to human post-editing.

accounting for the complexity of such operations as a function of linguistic properties of the text as it is done in the cognitive dimension.

As previously mentioned, the most intuitive and direct measure of post-editing effort is **post-editing time**. The time taken to post-edit can be used as a proxy for quality: segments that take longer to be post-edited are considered worse than segments that can be quickly corrected. Koponen et al. (2012) argue that post-editing time is the most effective way of measuring cognitive aspects of the post-editing task and relating them to the quality of the machine translations. Plitt and Masselot (2010) use post-editing time (more specifically, words per hour) to measure the productivity gain of post-editing machine translated text in a real scenario of translation workflow, instead of performing translation from scratch. The words per hour metric shows high variation among different annotators, although post-editing is consistently less time-consuming than translation from scratch. The authors also show that MT reduces keyboard time (time that the translator spent typing) by 70% and pause time (time that translator spent thinking, reading and searching for references) by 30%. Although post-editing time seems to be a good metric of post-editing effort, it can be inaccurate and difficult to achieve. Firstly, the post-editing time is a very noisy metric, since the translators can get distracted or take breaks while translating a sentence. Secondly, a high variation among different translators' post-editing time is expected, given that translators have different typing skills, translation experience and proficiency with the post-editing tool, among other aspects. In addition, post-editing time can encompass reading time, correction time and revision time, although the relationship among these factors is unclear.

Perceived post-editing effort is an alternative way of evaluating post-editing effort and it can capture cognitive aspects of post-editing. In this evaluation approach, humans are asked to give a score for the machine translated sentences according to a *likert* scale (Specia et al., 2011). This type of scores can be given with or without actual post-editing and they represent the humans belief on how difficult it would be (or it was) to fix the given machine translated sentences. In the first edition of WMT QE shared task in 2012 (Callison-Burch et al., 2012), the *likert* scale varied from 1 to 5, where:

- 1 - The MT output is incomprehensible, with little or no information transferred accurately. It cannot be edited, needs to be translated from scratch.
- 2 - About 50% to 70% of the MT output needs to be edited. It requires a significant editing effort in order to reach publishable level.
- 3 - About 25% to 50% of the MT output needs to be edited. It contains different errors and mistranslations that need to be corrected.

- 4 - About 10% to 25% of the MT output needs to be edited. It is generally clear and intelligible.

- 5 - The MT output is perfectly clear and intelligible. It is not necessarily a perfect translation, but requires little to no editing.

Alternatively, cognitive aspects of post-editing can be measured via **eye-tracking**. Eye-tracking techniques can measure fixation time (for how long the post-editor fixates on the text) or saccade time (movement of eyes). O'Brien (2011) measures fixation time and correlates it with General Text Matcher (GTM) (a similarity metric between the target and the reference sentences based on precision, recall and f-measure) (Turian, Shen, and Melamed, 2003). Fixation time shows correlation with GTM scores: low GTM scores shows high fixation time. In addition, post-editing pauses (extracted from keystroke logs) can also be viewed as an indirect measure of cognitive effort (Lacruz, Denkowski, and Lavie, 2014). Long pauses are associated with segments that demand more cognitive post-editing effort.

Blain et al. (2011) define Post-Editing Action (PEA) as a post-editing effort metric based on **linguistic information**. The difference from "mechanical" edits (e.g. insertion, deletion, substitution) is that PEA is related to linguistic changes. Instead of simple word edits, PEA has a "logical" aspect that aggregates several "mechanical" edits together, based on the type of change. For example, if changing a noun in a sentence would mean changing the gender of an adjective in the same sentence, PEA would count only one change (the noun) while "mechanical" edits would take into account both changes (noun and adjective). In order to automatically identify the PEAs, Blain et al. (2011) use TER as proxy for for edit distance. The authors claim that informing users with PEAs information would help in cases where the MT quality is already high and the task is to perform light reviews.

In general, although cognitive effort is an important dimension of the post-editing process, its measurement is usually expensive and unreliable. Perceived post-editing effort can be highly influenced by differences in the view of annotators and how accepting of MT they are. Eye-tracking techniques and post-editing pauses are not immune to cases where the post-editor is not focused on the task and start to digress (a similar problem is presented in measuring post-editing time). On the other hand, linguistic-based approaches are expensive to be performed manually and require accurate NLP tools in order to be performed automatically and are, therefore, highly language-dependent.

Finally, post-editing effort can also be evaluated indirectly, by using a metric that takes into account **edit operations** (technical effort). HTER is an example of such metrics. Proposed by Snover et al. (2006), in HTER post-edited machine translation are compared to the

original machine translation, by using TER. HTER then computes the minimum number of edits to transform the machine translation into the post-edited version. Although HTER is less time-consuming, it is still not a clear evaluation of post-editing effort to inform human translators. A human post-editor would probably focus on performing the task as fast as possible, instead of minimising the number of edits (Koehn, 2010). However, HTER is still widely used as an indirect measurement of post-edit effort (Bojar et al., 2013, 2014, 2015, 2016b).

Recent work has discussed the use of target-side **fuzzy match scores** in order to evaluate MT outputs (Parra Escartín and Arcedillo, 2015). Traditional Translation Memory (TM) fuzzy match scores are similarity scores between the sentence to be translated and the TM entries, ranging from 0% (no similar sentence was found in the TM) to 100% (an exact sentence was found in the TM). The translation industry normally operates discounts in costs according to the fuzzy match scores. The target-side fuzzy match scores aim to provide a similar metric for MT outputs. Parra Escartín and Arcedillo (2015) calculates such scores by comparing the machine translated sentences with their post-edited versions. Their results show that target-side fuzzy match scores show high Pearson's r correlation with BLEU and TER. Moreover, the authors claim that fuzzy match scores are more intuitive and widely used in the industry, which makes them easier to be assimilated by professional translators than automatic evaluation metrics widely used in academia.

2.3.2 End-user-based Evaluation

In contrast to post-editing cost, the aim of end-user-based approaches is to evaluate how reliable a machine translation is for the end-user. Such evaluation methodology is related to MT for assimilation purposes: the end-user needs to comprehend the machine translation and it can be achieved without perfect quality.

Reading Comprehension Tests

Reading comprehension tests can be used to evaluate MT in a scenario where the aim is to decide whether or not the machine translated text is comprehensible and encodes the main source information correctly. Such tests are given to humans and the correctness of the answers is used as the quality scores. Although this approach can be biased by the knowledge of the test taker, it is a fast and easy task to perform.

The usefulness of reading comprehension tests for MT evaluation has been addressed in previous work. Tomita et al. (1993) use Test of English as a Foreign Language (TOEFL)

reading comprehension tests to evaluate English into Japanese machine translations. Texts were machine translated by three different MT systems, the questions were manually translated and native speakers of Japanese answered the questions. Results show that reading comprehension tests are able to distinguish the quality of the three MT systems. Fuji (1999) uses reading comprehension tests for an official examination of English as a second language, designed especially for Japanese native speakers. The author focuses on evaluating seven MT outputs, regarding their quality. Reading comprehension tests are used to evaluate the "informativeness" of the MT outputs, but the results only show significant differences among two systems. Therefore, the majority of the MT systems are not distinguished in terms of "informativeness". Fuji et al. (2001) evaluate the "usefulness" of MT by using Test Of English for International Communication (TOEIC) reading comprehension tests. Again, the language pair was English-Japanese, but in this work the machine translated version was evaluated by native speakers in two scenarios: (i) only the machine translated version was shown and (ii) the source was also shown. Scenario (ii) showed better results than (i) for native speakers with low proficiency in English, for which the TOEIC scores improved and the time spent answering the questions decreased.

Later, Jones et al. (2005b) use the Defence Language Proficiency Test (DLPT) structure to evaluate the readability of Arabic-English MT texts. Their results show that subjects are slower at answering questions on the machine translated documents and that their accuracy is also inferior compared to human translated documents. Jones et al. (2005a) also use DLPT-style questions, aiming to find which level of Arabic reading comprehension a native speaker of English could achieve by reading a machine translated document. Their results show that MT texts lead to an intermediate level of performance by English native speakers.

More recently, Berka, Černý, and Bojar (2011) use a quiz-based approach for MT evaluation. They collected a set of texts in English, created yes/no questions in Czech about these texts and machine translated the English texts by using four different MT systems. The texts consist of small paragraphs (one to three sentences) from various domains (news, directions descriptions, meeting and quizzes). Their results show that outputs produced by different MT systems lead to different accuracy in the annotators' answers.

Scarton and Specia (2016) present a corpus for reading comprehension evaluation of MT systems. Such corpus was created from the Corpus of Reading Comprehension Exercises in German (CREG) (Ott, Ziai, and Meurers, 2012), by machine translating the German documents into English. The reading comprehension questions were also machine translated and then post-edited by a professional translator. Fluent native speakers of English were

then asked to answer the reading comprehension questions about the machine translated documents. More details about the use of this corpus in this thesis are presented in Chapter 6.

Eye Tracking and Task-based Questionnaires

Eye tracking techniques can also be applied in a scenario of evaluation for dissemination purposes. Doherty and O'Brien (2009) evaluate machine translated sentences by using eye tracking while native speakers of the target language (French) read the sentences. Results of gaze time, fixation count and average fixation duration are compared with a manual analysis of the sentences (sentences were previously classified as "good" or "bad"). Moderate Spearman's ρ correlation ranks are found between the manual analysis and gaze time and fixation count, meaning that long gaze times and high number of fixation counts correlate with "bad" sentences while "good" sentences lead to short gaze times and low number of fixation counts. Doherty, O'Brien, and Carl (2010) extend the work of Doherty and O'Brien (2009) and compare the eye tracking measurements with BLEU scores at sentence level. The authors find a trend where "bad" sentences present long gaze times, high number of fixation counts and low BLEU scores, whilst "good" sentences present short gaze times, low number of fixation and higher BLEU scores.

Stymne et al. (2012) combine eye tracking and error analysis in order to evaluate the quality of three different MT systems and a human translation from English into Swedish. Reading comprehension questions and *likert* scales for fluency, comprehension, confidence of correct answers and estimated errors are also used in the evaluation performed by native speakers of Swedish. Moderate Pearson's r correlation is shown between eye tracking gaze time and fixation time and the other measurements, although it does not happen among all MT systems.

Doherty and O'Brien (2014) use eye tracking and post-task questionnaires to evaluate the usability of machine translated texts in helping users to perform a task (use of a file storage service). The original instructions were in English, while translations were provided for French, German, Spanish and Japanese. Results show that the original English documents require significative less cognitive effort (in terms of total task time, fixation counts and average fixation time) than the Japanese documents. The other languages do not show significative results when compared to English. Castilho et al. (2014) also use eye tracking and post-task questionnaires, although their aim is to evaluate whether or not post-editing increases the usability of machine translated technical instructions performed by end-users. The language pair was English - Brazilian Portuguese and the users were asked to perform some tasks using a security software for personal computers. Post-edited versions appear

to be significantly more usable than machine translations. Castilho and O'Brien (2016) follow the work of Castilho et al. (2014) and also evaluate the usability of post-editing for English-German instructions in how to use a spreadsheet application. No significant difference is shown between machine translated documents and post-edited documents, in terms of cognitive effort measured by eye tracking. However, users needed less time to perform the tasks with the post-edited versions and showed more satisfaction with this kind of text than with machine translated versions.

Klerke et al. (2015) evaluate the performance of humans in solving puzzles by using the original source version (English), a simplified version of the source, a human translation of the original source into Danish and machine translations of the original and simplified source into Danish. They use eye-tracking, time, comprehension and comparison tests to gather information about the difficulty of the task. They also compare their measurements with BLEU and task performance. Humans showed comparable performance for both human translations and simplified versions of the source. This indicates that simplified versions of the source help as much as human translations. The machine translated versions of simplified texts improve reading processing over the original text and the machine translation of the original text, although the performance of the participants was better when using the original source documents rather the machine translations. It seemed that simplifying documents before machine translating it can produce better results than machine translating the original. In terms of eye-tracking measurements: (i) humans spent more time on machine translated versions (either simplified or not) than on original documents; (ii) fixations were higher for machine translated versions and lower for human translations; and (iii) more regressions (from a word read to a previous point in the text) were observed in machine translated versions. BLEU does not show correlations with task efficiency, while time and fixations metrics show significant Pearson's r correlation scores.

Finally, Sajjad et al. (2016) propose the use of eye tracking measurements to predict the perceived quality of machine translated sentences from English into Spanish. They selected the best and worst machine translations of 60 sentences from the WMT12 evaluation shared task data. Although this work is not directly related to the use of eye tracking for task-based approaches, it shows some important advances in the use of eye tracking for MT evaluation purposes. Eye-tracking measurements refer to reading processing: progression (jumps forward) and regression (jumps backwards) along the words in the sentence, jump distances, jumps between machine translation and reference, fixation time and lexicalised information of words being read. The authors build regression models using the eye-tracking measurements as features. They show that eye-tracking measurements from the target and

lexicalised features are the best in predicting the ranks. Moreover, when combined with BLEU, such eye-tracking measurements show higher Kendall's tau correlation with human ranks than BLEU alone.

2.4 Quality Estimation

This section revisits the concepts of QE and presents related work that was the basis for the research reported herein. In Section 2.4.1 the background and early work are presented and discussed. Section 2.4.2 contains state-of-the-art work on document-level QE. Section 2.4.3 presents a literature review of previous work focusing on linguistic features for QE. This section is included because we explore discourse-level linguistic information, and therefore, work on the use of linguistic features for QE is important for this thesis. Section 2.4.4 presents a state-of-the-art toolkit for QE, highlighting our contributions to it during the development of this research. Finally, Section 2.5 summarises the content of this chapter.

2.4.1 Introduction

The task of QE consists in predicting the quality of the outputs of MT systems without the use of reference translations. Work in QE began in the early 2000s and the focus was on Confidence Estimation (CE) of words (Gandrabur and Foster, 2003; Ueffing, Macherey, and Ney, 2003; Blatz et al., 2004; Ueffing and Ney, 2005; Ueffing, 2006; Ueffing and Ney, 2007) and sentences (Akiba et al., 2004; Blatz et al., 2004; Quirk, 2004).

CE can be viewed as a sub area of QE that focuses on estimating the confidence of a system in translating a word, a phrase or a sentence. According to Blatz et al. (2004), two main approaches were said to be possible for CE. Either the CE system is completely connected to the MT system (using the probabilities derived from the system as the confidence probabilities) or it is independent (using features of words/sentences, with the best features still related to the MT system under evaluation - e.g. from n-best lists). On the other hand, QE refers to a broader scenario in which it is possible to predict the quality of MT systems outputs, disregarding the systems themselves. Therefore, QE encompasses all strategies:

(i) CE approaches, focusing on the use of features dependent on MT systems;

(ii) approaches that only use features from words, sentences or documents (disregarding the MT systems information);

(iii) a combination of both.

It is worth mentioning that features that are not dependent on the MT system can make use of external resources (e.g. Language Models (LMs)). These features are important for cases where the system is unknown or there is no information available about its internal properties. An example is the use of QE for *gisting* in which the MT system is a black-box for the end users. However, it is worth emphasising that the use of MT systems information can be valuable in cases where it is available and, therefore, a model built following (iii) can be more effective.

More recent work on QE explores features that are dependent and independent from MT systems to predict the quality of words or sentences (Specia et al., 2009a,b; He et al., 2010; Specia and Farzindar, 2010; Specia et al., 2011). Different language pairs have been explored and a variety of quality labels were considered (e.g. BLEU-style metrics, HTER scores, *likert* scores). Regression has also been established as the best supervised ML technique to build prediction models for the task of sentence-level QE. The preference for regression occurred as a result of the focus on fitting a model to continuous labels (e.g. HTER) or an ordinal value (e.g. *likert* scores).

Shared tasks on QE have been organised since 2012, as part of the WMT (Callison-Burch et al., 2012; Bojar et al., 2013, 2014, 2015, 2016b). Word and sentence-level QE have been explored with a variety of data, quality labels and ML approaches. The latest two editions of WMT also included subtasks on paragraph and document-level predictions.

Granularity Levels of QE

Sentence-level QE is the most popular variety of the prediction and feature extraction levels. One of the reasons for this is the direct applicability in the industry, e.g., distinguishing segments to be post-edited from the ones that would take more time to post-edit than to translate from scratch (Specia et al., 2011). Moreover, it is more natural for humans to give a quality score for sentences (such as *likert* scores) and also to post-edit sentences rather than words or entire documents. This task is often addressed as a regression problem, with different algorithms explored (e.g. Support Vector Machines (SVM) (Cortes and Vapnik, 1995), Gaussian Process (GP) (Rasmussen and Williams, 2006)) and a rich range of features employed, including linguistically motivated ones.

In word-level QE (Blatz et al., 2004; Luong, 2014; Ueffing and Ney, 2005) each word receives a label that represents its quality. This level of prediction is useful, for example, to highlight errors in a post-editing/revision workflow. The word-level QE subtask at the WMT has been organised annually since 2013 and received more attention during the WMT15 edition, in which participants could outperform the baseline system. Classification is used

to categorise words into "GOOD" or "BAD" classes (although multi-class problems have also been explored in the WMT12 and WMT13 shared tasks). More recently, research on phrase-level QE is emerging (Logacheva and Specia, 2015; Blain, Logacheva, and Specia, 2016). For this task, groups of consecutive words (phrases) are labelled according to their quality. In WMT16 a first version of this task was organised where phrases were extracted from word-level annotations. Summarising, a group of consecutive words was considered a phrase if all words share the same quality label ("GOOD" or "BAD"). Therefore, the phrase-level QE task was also modelled as binary classification.

Document-level QE (Soricut and Echihabi, 2010; Scarton and Specia, 2014a; Scarton, 2015), the focus of this thesis, is the least explored of the three levels. An example of the usefulness of document-level QE is *gisting*, mainly when the end-user does not know the source language (more details are presented in Section 2.4.2). In theWMT15, we made the first effort towards a document-level QE, by organising a subtask in paragraph-level features and predictions. Only a few teams submitted systems for the task and METEOR was used as quality label. Therefore, only a few conclusions could be drawn from the first edition of this subtask (presented in Section 2.4.2). Winning submissions addressed the task as a regression problem, using similar techniques to sentence-level QE. In the latest edition of WMT (2016), we organised a document-level QE task, that used document-aware quality labels for prediction (the labels were extracted from a two-stage post-editing method). Again, only a few teams submitted systems and the winning submissions explored GP models, with a kernel combination of handcrafted and word embeddings features (more details about the quality label used are presented in Chapter 6).

General Framework for QE

The general framework for QE is illustrated in Figures 2.1 and 2.2. In Figure 2.1, labelled words, sentences or documents from the source and the target languages are used as inputs for extracting features. Information from the MT system can be also used to extract features for QE. These features can be simple counts (e.g. length of sentences in the source or target texts), explore linguistic information (e.g. percentage of nouns in the source or target texts); or include data from the MT system (e.g. an n-best list used to build a LM and compute n-gram log-probabilities of target words). The features extracted are then used as the input to train a QE model. This training phase can use supervised ML techniques, such as regression. In this case, a training set with quality labels is provided. These quality labels are the scores that the QE model will learn to predict. The quality labels can be *likert* scores, HTER, BLEU,

to cite some widely used examples. The ML algorithm can vary, where Support Vector Machines (SVM) and Gaussian Process (GP) are some examples.

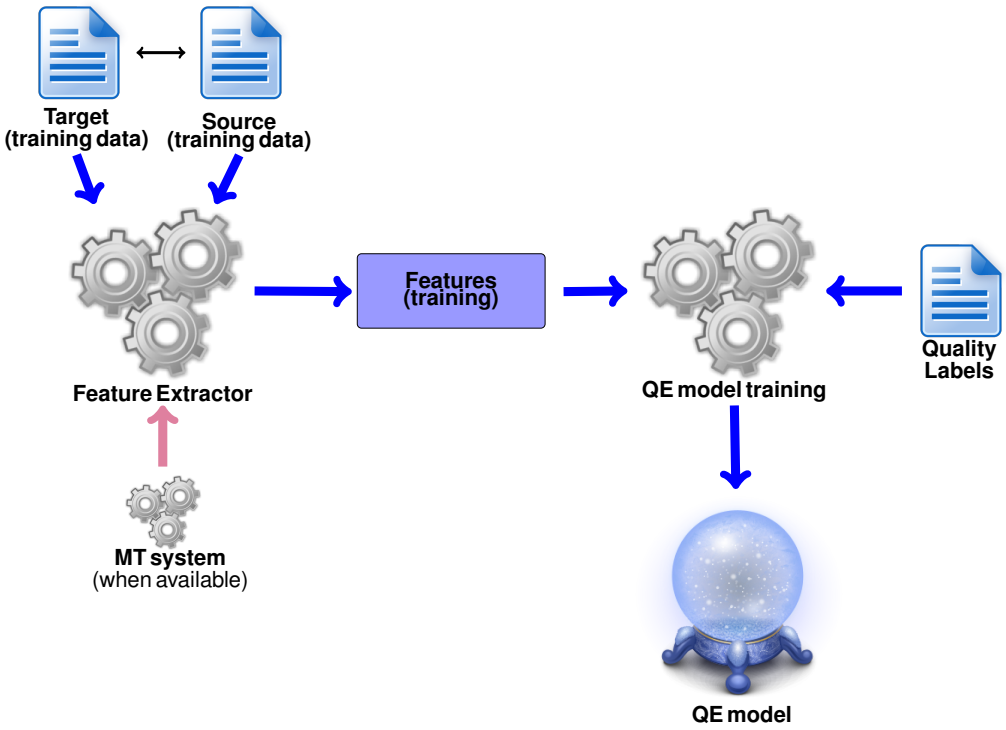

Fig. 2.1 General framework of Quality Estimation: training stage

Figure 2.2 illustrates how unseen data is labelled by a QE model. Features are extracted from the unseen data in the same way they were extracted from the training data. The extracted features are then given to the QE model trained previously. Based on what the model learned from the training data, it will predict scores for the unseen data.

2.4.2 Document-level Prediction

Assessing quality beyond sentence-level is important for different scenarios:

- *gisting* - mainly when the end user does not understand the source language;
- completely automatic solutions (e.g.: automatic translation of news, automatic translation of reviews, etc.).

Document-level QE is a challenging task mainly because of three reasons. Firstly, it is not as straightforward to assess documents as it is to assess sentences or words. Humans

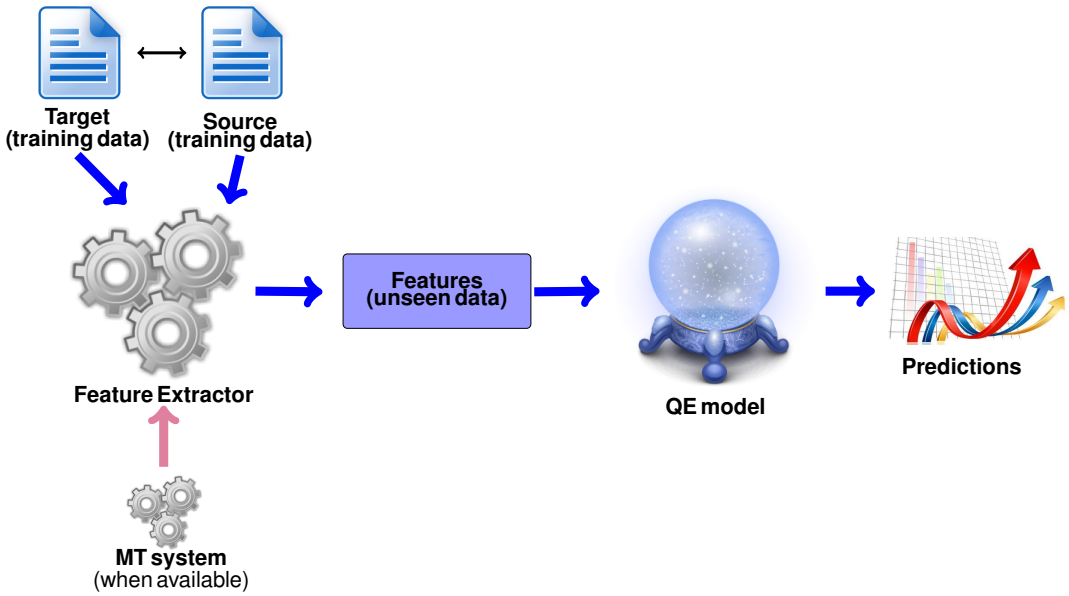

Fig. 2.2 General framework of Quality Estimation: predicting quality of unseen data

are able to give quality scores (such as adequacy and fluency scores, varying from 1 to 5) for words and even sentences, but not for documents (Scarton et al., 2015). Classifying large units of texts into fixed scores at document level is difficult because small problems at sentence and word level interfere in the human judgements. Secondly, there is very little parallel data with document markups available. Whilst thousands of sentences and words can be found easily, documents are limited. Finally, as a less studied level, adequate features still need to be explored for document-level QE.

Regarding document-level labels, previous work use BLEU-style metrics as quality labels for prediction (Soricut and Echihabi, 2010; Scarton and Specia, 2014a; Scarton, 2015; Bojar et al., 2015). The problem with these metrics is that they were not developed for document assessment in absolute terms, hence they are more useful for system comparisons. As a result, different documents machine translated by the same MT system tend to be scored similarly, even though they can actually have different quality. Moreover, it is also not ideal to consider the simple aggregation of sentence-level scores as the document score. Sentences can have different relevance to the document. Sentences that are more important for the document understanding are more problematic if incorrectly translated than sentences that play a less important role in the document (Turchi, Steinberger, and Specia, 2012). Additionally, sentences can score very badly when tested in isolation, although they can be considered acceptable when put in the document context.

In terms of data, as mentioned before, the majority of MT systems translate documents sentence by sentence, disregarding document information. Therefore, for practical reasons, the majority of the parallel corpora available for MT does not have document markups. Such corpora are collections of random sentences, making their use impossible for document-level approaches. Moreover, available parallel corpora with document-level mark-ups have hundreds of documents, while to train a model for QE it would be preferable to have more data points.

Several features have been proposed for QE at sentence-level. Many of them can be directly adapted for document-level (e.g. the number of words in source and target sentences can be extended to number of words in source and target documents). However, other features that better explore the document as a whole or discourse-related phenomena can be more informative. An issue in this case is the lack of reliable NLP tools for extracting discourse information. Although there have been recent advances in this area for English, other languages are still not covered. Therefore, features for QE that address discourse are limited to source or target language only and when translating from or into English. Initiatives to learn discourse relations without supervision (or at least to learn relations beyond sentence boundaries) are promising, as they are language independent. In Section 2.4.3, we present work addressing linguistic features for QE, showing the motivation for using discourse information based on the improvements achieved by using linguistic information for word and sentence-level QE.

Previous Work on Document-level QE

Apart from our own work that we present in this thesis, only a few other studies address document-level QE. Soricut and Echihabi (2010) explore document-level QE prediction to rank documents translated by a given MT system, predicting BLEU scores. Features include text-based, language model-based, pseudo-reference-based, example-based and training-data-based. Pseudo-reference features are BLEU scores based on pseudo-references from an off-the-shelf MT system, for both the target and the source languages. The authors discuss the importance of the pseudo-references being generated by MT system(s) which are as different as possible from the MT system of interest, and preferably of much better quality. This should ensure that string similarity features (like BLEU) indicate more than simple consensus between two similar MT systems, which would produce the same (possibly bad quality) translations. While promising results are reported for ranking of translations for different source documents, the results for predicting absolute scores proved inconclusive. For two out of four domains, the prediction model only slightly improves over a baseline

where the average BLEU score of the training documents is assigned to all test documents. In other words, most documents have similar BLEU scores, and therefore the training mean is a hard baseline to beat.

Soricut and Narsale (2012) also consider document-level prediction for ranking, proposing the aggregation of sentence-level features for document-level BLEU prediction. The authors claim that pseudo-reference features (based on BLEU) are the most powerful in the framework.

Scarton and Specia (2014a) explore discourse features (based on LSA cohesion) and pseudo-references for document-level QE. Although the pseudo-reference features were the best among the baseline and discourse features, LSA features also showed improvements over the baseline. Scarton and Specia (2015) study the impact of discourse phenomena in machine translation, by measuring the correlation between discourse features and HTER scores (more details on this work are presented in Chapters 4 and 5). Scarton et al. (2015) discuss the use of automatic evaluation metrics as quality labels for QE and propose a new human-targeted method of post-editing, focusing on isolating document-level issues from more fine-grained issues (see Chapter 6). Scarton and Specia (2016) present a new corpus with reading comprehension scores, annotated by humans, also aiming at new document-level labels (see Chapter 6).

Two teams participated at the WMT15 document-level QE task. Scarton, Tan, and Specia (2015) describe two systems. The first system uses QUEST++ features and discourse information (from discourse parsers and LSA - see Chapter 4) and a feature selection approach based on Random Forests and backward feature selection. The second system performs an exhaustive search on the official baseline features.[3] This system achieved performance similar to the first one, by only using three features out of the 17 baseline features. The system of Biçici, Liu, and Way (2015) applies Referential Translation Machines (RTM) for document-level QE, performing as well as the top system. RTMs (Biçici, 2013; Biçici and Way, 2014) explore features such as distributional similarity, the closeness between test instances to the training data, and the presence of acts of translation. The authors used Support Vector Regression (SVR) to train the models with feature selection.

At WMT16 two teams participated in the shared task of document-level QE. Scarton et al. (2016) describe two different approaches. The first uses word embeddings as features for building a QE model using GPs. Baseline features are combined with word embeddings by using different kernels into the GP model. The second system uses discourse information combined with baseline features for building SVR models for document-level QE. Besides

[3]This system was a joint work with Liling Tan from Saarland University, Saarbrücken, Germany.

the discourse features proposed by Scarton, Tan, and Specia (2015), the authors also use features from a entity graph-based model (Sim Smith, Aziz, and Specia, 2016a) to measure the coherence of the source and target documents (see Chapters 4 and 5 for more details about these systems). The second team use RTMs as previously used by Biçici, Liu, and Way (2015) (Biçici, 2016).

2.4.3 Previous Work on Linguistic Features for QE

Linguistic Features for QE at Word Level

Xiong, Zhang, and Li (2010) use linguistic features (combined with MT system-based features) for error detection in Chinese-English translations. They cover lexical (combination of words and part-of-speech tags) and syntactic features (whether or not a word was ignored by the Link Grammar parser, because the parser could not interpret it). They consider the task as a binary classification problem (a given word is correct or not). The result of the combination of linguistic features with system-based features outperforms the baseline and the results of only using lexical or syntactic features outperform the results of only using MT system-based features.

Bach, Huang, and Al-Onaizan (2011) also apply word-level quality estimation integrating linguistic features from a part-of-speech tagger. Features are calculated at word level, although they extend the model to sentence-level. Like Xiong, Zhang, and Li (2010), Bach, Huang, and Al-Onaizan (2011) address the task as a classification problem. The use of part-of-speech features in the source segment led to the best result. Later Luong (2014) proposes new syntactic features for word-level QE and combined them with features proposed in previous work. The new syntactic features are also extracted from the Link Grammar parser, such as the constituent label of the words and the word depth in the syntactic tree. However, the authors do not report on the impact of the new syntactic features separately.

Martins et al. (2016) use syntactic features, combined with unigram and bigram features for building the winning system for the word-level QE shared task at WMT16. The syntactic features were extracted by using the TurboParser (Martins, Almeida, and Smith, 2013). Such features encompass information about dependency relations and syntactic head of target words. The authors show that syntactic features lead to improvements over the use of unigrams and bigrams features.

In summary, for word-level QE, state-of-the-art systems use linguistic information as features for the task. Features that use syntactic knowledge have shown promising results. Document-wide information, however, has never been explored mainly due to dataset con-

straints (the datasets for word-level QE tasks are usually composed of randomised sentences). Therefore, there is also room for using the findings of this thesis for word-level QE.

Syntactic Features for QE

Specia et al. (2011) predict the adequacy of Arabic-English translations at sentence level, using a classification approach (predicting *likert* scores) and a regression approach (predicting METEOR). They group the features used in four classes: (i) source complexity features (e.g. average source word length); (ii) target fluency features (e.g. target language model); (iii) adequacy features (e.g ratio of percentage of nouns in the source and target sentences); and (iv) confidence features (e.g. SMT model score). Linguistic features are used for the first three categories and cover different linguistic levels. An example of syntactic features is the absolute difference between the depth of the syntactic trees of the source and target sentences. They do not evaluate directly the impact of only using linguistic features, however, 6 out of 12 adequacy features contained linguistic information.

Avramidis et al. (2011) consider syntactic features for ranking German-English SMT systems. The syntactic information is generated using a Probabilistic Context-Free Grammar (PCFG) parser on the target and source sentences. The best results are obtained when syntactic features are used. Similarly, Almaghout and Specia (2013) use Combinatory Categorical Grammar (CCG) in order to extract features for QE. They apply these features to the output of French-English and Arabic-English systems. The CCG features outperform the PCFG features of Avramidis et al. (2011) when no other features are included.

Hardmeier (2011) and Hardmeier, Nivre, and Tiedemann (2012) apply syntactic tree kernels (Moschitti, 2006) to QE at sentence level for English-Spanish and English-Swedish machine translations. The syntactic information is encoded into the model as tree kernels, which measure the similarity of sub-syntactic trees. The use of tree kernels leads to improvements over a strong baseline. Beck et al. (2015) also explore syntactic tree kernels to QE for French-English and English-Spanish language pairs. Although the focus of the paper is in hyperparameter optimisation via Bayesian methods (in this case, GP), the authors report on improvements for QE over the method studied by Hardmeier (2011).

Kozlova, Shmatova, and Frolov (2016) explore morphosyntactic, LM-based, pseudo-reference and baseline features in order to build the best system for sentence-level QE at WMT16. The syntactic features were extracted by a dependency parser. Syntactic features from source and target sentences are: tree width, maximum tree depth, average depth of the tree and proportion of internal nodes of the tree. Source sentence only syntactic features are: number of relative clause and number of attributive clauses. Information from POS-tags

were also extracted for source and target sentences. Although the winning system was a combination of syntactic features and all other features, the authors show that the combination of baseline and syntactic features outperform the baseline system. However, when performing feature selection, only one out of ten features is syntactic. Finally, Sagemo and Stymne (2016) also explore syntactic features for the sentence-level shared task at WMT16. Their syntactic features are similar to the one used in Avramidis et al. (2011), encompassing information extracted from a PCFG parser. Baseline results outperform by the combination of baseline features and syntactic information.

The use of syntactic features outperforms strong baselines for the task of sentence-level QE. In fact, it is expected that syntactic information should play an important role in sentence-level evaluation, since several problems with the output of MT systems are at the grammatical level. Systems built with syntactic knowledge still perform best in shared task competitions.

Semantic Features for QE

Pighin and Màrquez (2011) use Semantic Role Labelling (SRL) to rank English-Spanish SMT outputs. They automatically annotated the SRLs in the source side and project them into the target side, by using the word alignments of the SMT system. The evaluation is done by considering the human assessments available for the WMT13 2007-2010 corpora and the dataset described in Specia and Farzindar (2010). They train the model in Europarl data, therefore, they separate the data into in-domain data (Europarl) and out-of-domain data (news). The results of using SRL features are better than the baseline for the in-domain data. Some results of out-of-domain data are comparable to in-domain data when SRL features are applied. Specia et al. (2011) also explore semantic features, such as the difference in the number of person/location/organization named entities in source and target sentences. As mentioned before, the authors do not present an analysis of the linguistic features used, although 50% of the best adequacy features were linguistic features.

It is still not clear what the contribution of semantic knowledge for sentence-level QE is. For sentence-level semantics (such as SRL) the lack of reliable resources is problematic for achieving competitive results.

Analysis of the Impact of Linguistic Features in QE

Felice and Specia (2012) introduce several linguistic features for English-Spanish QE. They cover three linguistic levels: lexical (e.g. percentage of nouns in the sentence), syntactic (e.g. width and depth of constituency trees) and semantic (e.g. number of named entities).

Although the combination of linguistic features does not improve the results, an analysis of the contribution of features shows that linguistic features appeared among the top five. The application of a feature selection method result in a set 37 features, out of which 15 were linguistic. This selected set leads to improved results.

In general, previous work on linguistic features for sentence-level QE explore mainly lexical and syntanctic features. Results are promising for the majority of the scenarios, although it is clear that more work needs to be done in the case of semantic features. As for word-level, document-wide information has not been explored so far for sentence-level QE. Again, constraints on datasets are the main reason for such a lack of studies.

2.4.4 QUEST++: a Toolkit for QE

Toolkits for feature extraction and creation of QE models are available. The most widely is QUEST++[4] (Specia, Paetzold, and Scarton, 2015). QUEST++ is the newest version of QUEST (Specia et al., 2013), with support for word, sentence and document-level feature extraction and prediction. It has two main modules:

- **Feature Extractor module**: extracts several features from source and target texts and also from MT systems (when available). This module is implemented in Java;
- **Machine Learning module**: provides wrappers to `scikit-learn`[5] ML algorithms for training and applying QE models. This module is programmed in Python.

For sentence and document-level QE, 17 features are commonly used as a baseline[6], including as of final baseline at the WMT shared tasks. These are basic features that count punctuation marks, tokens and *n*-grams, compute target and source language model perplexities and the number of possible translations per word in the source text.

Figure 2.3 shows the QUEST++ structure with the two modules. Our contribution is at the document-level feature extraction (part highlighted in Figure 2.3). The input files are sequences of paths to source and target documents. These documents are then processed individually. Since some features are an aggregation of feature values for sentences, the document itself is stored as a set of sentences.

We have implemented 78 features that encompass basic counts, LM probabilities, syntactic information and lexical cohesion (see Appendix A). The same algorithms used to train

[4]http://www.quest.dcs.shef.ac.uk
[5]http://scikit-learn.org/stable/
[6]http://www.quest.dcs.shef.ac.uk/quest_files/features_blackbox_baseline_17

sentence-level QE models are used to train document-level models, therefore no changes were made in this module for document level.

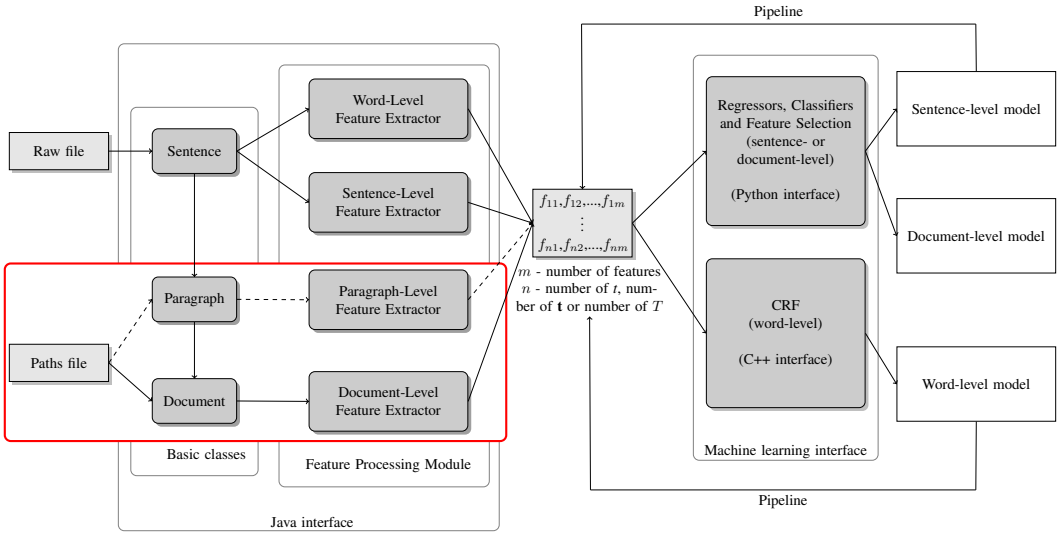

Fig. 2.3 QUEST++ framework structure

ASIYA[7] (Giménez and Màrquez, 2010) is an alternative framework for feature extraction at sentence and document level. ASIYA contains 24 features that can be extracted for sentence or document levels, whilst QUEST++ has 69 features that can be used for both granularity levels. Moreover, QUEST++ implements 9 features that are only document-level. Finally, QUEST++ also makes available a ML pipeline for training QE systems at document level. Marmot[8] and WCE-LIG[9] (Servan et al., 2011) are alternatives for word-level feature extraction and prediction. Qualitative[10] (Avramidis, 2016) is an alternative for sentence-level QE.

2.5 Discussion

In this chapter we presented the evaluation techniques used in MT. Although human evaluation (Section 2.1) is considered the most reliable type of MT assessment, this type of evaluation is time-consuming and impractical for *gisting* also, depending on the type of

[7] http://asiya.cs.upc.edu/
[8] http://qe-team.github.io/marmot/
[9] https://github.com/besacier/WCE-LIG
[10] https://github.com/lefterav/qualitative

evaluation, the variation between human assessments can be an issue for the reliability of the evaluation.

Automatic metrics for MT evaluation have been proposed to overcome some of the issues of human evaluation (Section 2.2). These metrics are considerably faster to compute and are useful for comparing different MT systems. Nevertheless, automatic metrics also have drawbacks such as the need of human references, low intuitiveness and the difficulty of adapting them for use in translation workflows.

Task-based approaches for MT evaluation are an alternative that are more intuitive than automatic evaluation metrics (Section 2.3). Such an evaluation type aims to assess MT considering the purpose it will be used for. For example, if the purpose is to improve the productivity of human translation workflows, post-editing effort information of machine translated texts can inform the translators about the quality of the MT they will be translating. Similarly, if the purpose is *gisting*, reading comprehension questions about the machine translated document can be helpful in indirectly assessing the quality of the MT. Such approaches are promising since they take into account the practical purpose of MT.

QE of MT uses annotated data points in order to train ML models (Section 2.4). The quality annotation can be any of the previous types, including human evaluation, automatic metrics or task-based approaches. However, identifying the right quality label to be used for training the QE models is an open challenge.

Document-level QE is a level much less explored than sentence and word-level QE. For scenarios such as *gisting*, document-level QE is desirable. Therefore, studying such a level is necessary in the QE field.

Linguistic features have been shown promising results for QE at word and sentence levels, improving over baseline results in the majority of the cases. However, the extraction of such features is not straightforward and the use of tools and language-dependent resources are needed. Moreover, no previous work has explored discourse features for QE. Finally, although document-level QE can benefit from syntactic and semantic features, document-wide linguistic information (such as discourse) has not yet been explored.

In the next chapter, we present a literature review on discourse processing in NLP focusing on work done for MT and MT evaluation. We also show the relation between each level of discourse processing and the features for document-level QE proposed by us in Chapter 4.

Chapter 3

Discourse Processing

Discourse Processing is an NLP area that aims to define the connections among textual parts that form a document. AS, TS and Readability Assessment (RA) are some examples of NLP tasks where discourse processing has already been applied. In MT, there have also been initiatives to integrate discourse information into different parts of the MT framework.

As mentioned before, traditional MT systems translate texts sentence-by-sentence disregarding document-wide information. This is the norm in SMT, where the core technique for decoding (dynamic programming) assumes sentence independence. An exception is Docent (Hardmeier, 2012), a document-level decoder for SMT approaches, based in local search to improve over a draft translation of the entire document. For each search state, Docent has access to the complete translation of the document. Small modifications are made in the search state aiming for a better translation. Although decoding is not the focus of this thesis, it is worth mentioning that Docent enables the use of discourse information into the decoding phase of SMT.

In this chapter, we define the discourse phenomena that are explored in this thesis (Section 3.1), as well as the work related to the use of discourse information for improving MT systems and for MT evaluation. We follow the definitions of Stede (2011) of cohesion and coherence and also the division of discourse phenomena in three levels: "large units of discourse and topics", "coreference resolution" and "small units of discourse". Although discourse processing encompasses a broad range of phenomena, an exhaustive review of all of them is out of the scope of this thesis. We cover only theories and techniques applied in the experiments we have performed.

Section 3.2 presents the theory behind "large units of discourse and topics", focusing on topic modelling, word embeddings and lexical cohesion. Such approaches consider the

document as a whole and refer to the cohesion of the document. Related work on using such techniques for MT is also discussed.

Section 3.3 shows the concept of "coreference resolution" that refers to the coherence of the document. Anaphora resolution and local coherence are the topics discussed and related work for MT is also shown.

Finally, in Section 3.4, "small units of discourse" is discussed, including discourse connectives and RST studies. These discourse units are responsible for the logical connections between parts of the text that form a coherent document. MT related work is also presented.

The reason we present work on the use of discourse to improve MT (instead of only showing work on evaluation and QE of MT) is that this work gave us important background for developing our discourse features (presented in Chapter 4). This literature review is not exhaustive and has the purpose of posing a discussion about the challenges faced in exploring discourse in MT.

3.1 Discourse Processing Background

A group of sentences generated by one or more people in order to share information is called a **discourse** (Ramsay, 2004). Each sentence in a discourse needs to be interpreted in the contexts of all sentences involved. Discourses can be from different sizes and, in some cases, the links are established among discourses and not among sentences (e.g. paragraphs, chapters, etc). According to Ramsay (2004), whether the discourse is a text produced by only one author or it is a dialogue among several speakers does not impact the processing of the information conveyed. In both scenarios, both reader and writer (speaker and hearer) need to share the same beliefs and previous background in order to establish an effective communication.

There are two well-know concepts related to discourse: **coherence** and **cohesion**. According to Stede (2011), a text is **coherent** if it is well written around a topic and each sentence is interpreted based on the other sentences. In order to build a coherent text, sentences are connected by coreferences (e.g. connectives, anaphora) or by implicit relations given by semantics and pragmatics. **Cohesion** is related to local discourse information. Different from coherence, where the text should be interpreted in order to find coherent relations, cohesion can be identified in the surface of the text, without a deep understanding of the text. We can interpret cohesion as the explicit signs of coherence (e.g. connectives, lexical chains).

To explain the diferences between cohesion and coherence, consider examples (i), (ii) and (iii).[1]

(i) "My favourite colour is **blue**. I like **it** because **it** is calming and **it** relaxes me. I often go outside in the summer and lie on the grass and look into the **clear sky** when I am **stressed**. **For this reason**, I'd have to say my favourite colour is **blue**."

(ii) "My favourite colour is **blue**. **Blue** sports cars go **very fast**. Driving **in this way** is dangerous and can cause many **car crashes**. I had a **car accident** once and **broke my leg**. I was very sad because I had to miss a holiday in Europe **because of the injury**."

(iii) "My favourite colour is blue. I'm calm and relaxed. In the summer I lie on the grass and look up."

Example (i) is coherent because the sentences make sense together (the person likes blue, because it is calm and relaxing). It is also cohesive, since the connections (highlighted) are made.

On the other hand, Example (ii) is a cohesive paragraph with no coherence. This paragraphs uses many cohesion devices and the sentences are well connected. However, the paragraph makes no sense: it contains a number of sentences describing different things connected together.

Finally, we could also find a coherent paragraph with no cohesion as in Example (iii). As in the first example, this paragraph shows a statement, a cause and an example, although it is more difficult to infer this logical meaning without the cohesive devices. However, one can still make logical connections mentally in order to understanding this paragraph. On the other hand, this paragraph is not cohesive since there are no connection among the sentences.

3.2 Large Units of Discourse and Topics

Large units of discourse and topics refer to genre. Documents from different genre can present different structure and word usage. For instance, scientific documents are different from news and literature. The differences are not only related to content, but also to how the information is structured and presented. Paragraphs in different positions of the documents play different roles. In scientific papers, for example, the introduction section is usually organised with the first paragraphs presenting background information and motivation and

[1]These examples were extracted from: http://gordonscruton.blogspot.ca/2011/08/what-is-cohesion-coherence-cambridge.html.

the last paragraphs summarise the work being presented. In contrast, news documents often present the background information as the last paragraph.

In addition, the genre will define the style of writing, which is directly related to lexical choices. For instance, news documents from a popular newspaper are not expected to use the same formal level that law agreements present. Humans adapt their writing style according to the genre expected.

Finally, lexical consistency can also be viewed as a phenomenon at this level. The art of keeping a document consistent in terms of word usage is associated to the topics presented in a document. For example, if a document addresses the Zika virus topic, it is very unlikely that information about football will appear.

3.2.1 Topic Modelling

Topic modelling refers to the group of algorithms that aim to identify topics in one or more documents and organise the information retrieved based on such topics (Blei, 2012). A topic can be seen as a theme that is explored on a document. Therefore, words such as "dogs" and "cat" are expected to correlate with the topic "pets", while they are not expected to correlate with the topic "golf war". Topics can be extracted using word frequency, Term Frequency - Inverse Document frequency (TF-IDF) and word probabilities as features for topic modelling, among others.

Latent Semantic Analysis (LSA) is a widely known topic modelling method (Landauer, Foltz, and Laham, 1998). This method is based on Singular Vector Decomposition (SVD) for dimensionality reduction. In SVD, a given matrix X can be decomposed into the product of three other matrices:

$$X = WSP^T \tag{3.1}$$

where W describes the original row entities as vectors of derived orthogonal factor values (a unary matrix with the left singular-vectors); S is a diagonal matrix containing scaling values (the singular values of X) and P (P^T is the transpose of P) is the same as W but for columns (a unary matrix with right singular-vectors). When these three matrices are multiplied, the exact X matrix is recovered. The dimensionality reduction consists in reconstructing the X matrix by only using the highest values of the diagonal matrix S. A dimensionality reduction of order 2 will consider only the two highest values of S, while a dimensionality reduction of order 3 will consider only the three highest values and so on.

In LSA, the X matrix can be formed by $words \times sentences$, $words \times documents$, $sentences \times documents$, etc. In the case of $words \times sentence$, each cell contains the frequency of a given

word in a given sentence. Usually, before applying SVD in LSA, the X matrix is transformed wherefore each cell encapsulates information about a word's importance in a sentence and a word's importance in the domain in general. Landauer, Foltz, and Laham (1998) suggest the use of a TF-IDF transformation in order to achieve this goal.

Latent Dirichlet Allocation (LDA) is another method for topic modelling that consider a document as random mixtures of topics (Blei, Ng, and Jordan, 2003). It is a probabilistic model that assumes that the topic distribution have a Dirichlet prior.

For each document d in a corpus D, LDA assumes the following process:

1. Choose $N \sim \text{Poisson}(\xi)$ - (N is chosen from a Poisson distribution)
2. Choose $\theta \sim \text{Dir}(\alpha)$ - (θ is chosen from a Dirichlet distribution)
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$
   (b) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n

where N is the number of words in d, z_n is the n^{th} topic for the word w_n, θ is the topic distribution for d and α and β are the parameters of the Dirichlet prior related to the topic distribution per document and the word distribution per topic, respectively.

Related Work on MT

Zhao and Zing (2008) propose a bilingual latent variable approach to model cohesion in SMT for English-Chinese. Such approach extracts monolingual topic-specific word frequencies and lexical mappings from the parallel data. A bilingual topic model is used to define the probabilities of candidate translations in the decoding phase. The use of this approach outperformed a baseline in terms of BLEU.

Zhengxian, Yu, and Guodong (2010) explore LDA for Chinese-English SMT by extracting the LDA topics on the English side and including these topics as features in the decoding phase. Compared to a baseline, the authors found an improvement in terms of BLEU when using the topical information. Eidelman, Boyd-Graber, and Resnik (2012) also explore LDA for Chinese-English SMT decoding. They extract LDA topics using two approaches: (i) source text viewed as a document (global version) and (ii) sentence viewed as document (local version) (covering cases where no document mark-up is available). In the test set with document mark-ups, the decoder, using both global and local versions as features, outperforms a baseline in terms of BLEU.

Related Work on MT Evaluation and QE

Rubino et al. (2013) explore a LDA approach to create topic models in two different ways for QE at sentence level. First, the authors concatenate both sides of a large bilingual corpus at sentence level. Sentences of source and target language are treated as a single bag of words to generate a single topic model. Therefore, the topic model built contains the vocabulary of both languages. Second, they explore a polylingual model in which two monolingual topic models are built for each language. Features are extracted based on the distance between source and target languages at sentence level by using metrics such as Euclidean distance and the topic models generated. They experiment with Arabic-English, English-French and English-Spanish machine translations and report improved results by using topic models.

Scarton and Specia (2014a) use LSA information for extracting features for QE at document level, showing promising results. David (2015) also explores a latent semantic approach by using bilingual Latent Semantic Indexing (LSI) vectors for sentence-level QE. The author concatenate source and target sentences and built a bilingual vector with word frequency. After applying TF-IDF and dimensionality reduction, the vectors are used to compute similarities between source and target sentences in the test set. His system was submitted to WMT15 and was one of the winners of the competition.

In this thesis we focus on using LSA as a metric of local cohesion (Scarton and Specia, 2014a). We also extend the work of Rubino et al. (2013) and use LDA topics as features for document-level QE (Chapter 4).

3.2.2 Word Embeddings

A widely used framework, called *word2vec* (Mikolov et al., 2013a), is computationally more efficient than previous approaches on neural network language models. Such neural models have been used in order to extract word representations and proved better than LSA and LDA models for some tasks (Mikolov et al., 2013b). This framework encompasses two approaches: Continuous Bag-of-Words (CBOW) model and Continuous Skip-gram model. CBOW aggregates information from words in context in order to predict a given word. The context is usually composed by previous and posterior words (this context is often called *window*). The Skip-gram model, on the other hand, uses the information of a single word in order to predict words in a given context. The idea behind this model is to find a good word representation that is able to define good predictions for its context (Mikolov, Le, and Sutskever, 2013).

Although word embeddings are not directly related to topic modelling approaches, they present an alternative semantic representation of words. Our assumption is that when word embeddings are combined at document level, they are able to encode information about the lexical cohesion of documents (Scarton et al., 2016).

Related Work on MT

Martínez-Garcia, Bonet, and Màrquez (2015) extend the work of Martínez-Garcia et al. (2014) and implement monolingual and bilingual word vector models into Docent. Word vectors are acquired from a large corpus using CBOW. Although their results do not show improvements over a SMT baseline, the use of word vector models improved over a Docent baseline.

Related Work on MT Evaluation and QE

Shah et al. (2015b) explore vectors extracted from a Continuous Space Language Model (CSLM) as features for sentence-level QE. CSLM implements a deep neural network architecture, with more layers than models like CBOW and skip-gram. Results show only slight improvements over a baseline when such features are used. In the same work, the authors present results of the use of CBOW vectors as features for word-level QE. They trained monolingual vectors and also a bilingual approach based on the work of Mikolov, Le, and Sutskever (2013). Results show that the use of bilingual vectors lead to better results (when compared to the use of monolingual data alone), although it did not outperform a strong baseline system. Finally, Shah et al. (2015a) explore CSLM for sentence-level QE in several datasets showing that the use of such features outperform the baseline in all datasets.

In this thesis we experiment with *word2vec* representations for document-level QE, by aggregating word vectors for the entire document (Chapter 4).

3.2.3 Lexical Cohesion

Lexical cohesion is a phenomenon of discourse related to reiteration and collocation (Halliday and Hasan, 1976). **Reiteration** happens when lexical items are repeated in order to keep the document cohesive. Synonyms, near-synonyms or superordinates can also be used to ensure lexical variety, without losing cohesion. For example Halliday and Hasan (1976, p. 278):

(iv) "There was a large **mushroom** growing near her, about the same height as herself; and, when she had looked under it, it occurred to her that she might as well look and see

what was on the top of it. She stretched herself up on tiptoe, and peeped over the edge of the **mushroom**, ..."

(v) "Accordingly... I took leave, and turned to the **ascent** of the peak. The **climb** is perfectly easy..."

(vi) "Then quickly rose Sir Bedivere, and ran,
And leaping down the ridges lightly, plung'd
Among the bulrush beds, and clutch'd the **sword**
And lightly wheel'd and threw it. The grand **brand**
Made light'nings in the splendour of the moon..."

(vii) "Henry's bought himself a new **Jaguar**. He practically lives in the **car**."

In Example (iv) the word *mushroom* is repeated in order to make the cohesive connection. In Example (v) *ascent* and *climb* are synonyms. An example of near-synonym is presented in (vi) with *sword* and **brand**. Finally, Example (vii) shows a case where a superordinate is used to make the reference: **car**, a more general lexical item, is used to refer back to *Jaguar*.

Collocations is the lexical cohesion form where lexical items are correlated because they frequently occur together. They are relations between lexical items that are not as strong as the relations presented by items in reiteration. Instead, these relations are defined by word meaning (e.g. antonyms - *like* and *hate*), ordered series (e.g. week days), unordered lexical sets (e.g. co-hyponyms - *chair* and *table*, both hyponyms of furniture). In summary, detecting collocations is a more challenging lexical cohesion phenomenon than reiteration, since the relations between the lexical items are weaker and, sometimes, very hard to automatically identify.

Related Work on MT

Carpuat (2009) explores the "one translation per discourse" hypothesis, based on the "one sense per discourse" hypothesis of the Word Sense Disambiguation (WSD) area (Gale, Church, and Yarowsky, 1992). This hypothesis says that each discourse should have the same translation within the entire document, in order to preserve consistency. She experiments with French-English human and machine translated texts and, in both cases, the lexical choices are considerably consistent. However, when comparing the machine translations with the reference translations, the majority of the cases where the references were consistent and the machine translations not, are, in fact, errors of the MT system. Therefore, the author proposes a post-processing stage to enforce "one translation per discourse" in the SMT outputs that

achieves slight improvements in BLEU, METEOR and National Institute of Standards and Technology (NIST) metrics. Ture, Oard, and Resnik (2012) also explore "one translation per discourse" hypothesis for Arabic-English and Chinese-English texts. They use a hierarchical phrase-based MT system based on Synchronous Context-Free Grammar (SCFG), which allows them to use the grammar rules for identifying the phrases that appear repeatedly in the source documents. The authors, then, use a forced decoding in order to implement the "one translation per discourse" hypothesis into the MT system. In forced decoding, the decoder uses rules from the SCFG learned from the translation pairs in order to find the derivations that transform the source sentence into the target sentence. Their approach outperforms a baseline systems (in terms of BLEU).

Carpuat and Simard (2012) make a detailed evaluation of lexical consistency in SMT for French-English, English-French and Chinese-English language pairs, using data from different sources and sizes. The authors found that SMT is as consistent in lexical choices as manually translated texts. In addition, the smallest dataset used presents higher consistency than its larger version (for Chinese-English), which can be explained by the size of the vocabulary being more restricted in a small corpus, favouring consistency. Although SMT can be generally consistent, the authors also show that the inconsistencies found in machine translated texts are more strongly correlated to MT errors than the inconsistencies found in human translated texts.

Xiao et al. (2011) explore lexical consistency to improve SMT at document level for Chinese-English. Ambiguous words are identified by their translations in the target language (if a word presents more than one translation in the target corpus it is considered ambiguous) and a set of translations candidates ($C(w)$) for each word (w) is defined. The authors employ two approaches: (i) post-editing: if t is a translations of the word w, but t does not appear in $C(w)$, t is replaced by a translation in $C(w)$ that guarantees consistency within the document; and (ii) re-decoding: in the first stage, the translation table is filtered eliminating any translation t that does not appear in $C(w)$ and then, on the second stage, the source sentences are decoded again using the filtered translation table. Their systems improve over the baseline by choosing the right translation for ambiguous words, in terms of error reduction. In terms of BLEU, only the re-decoding version outperforms the baseline.

Xiong, Ben, and Liu (2011) use lexical cohesion for document-level SMT in Chinese-English, by proposing three models to capture lexical cohesion: (i) direct rewards - rewards hypotheses that present lexical cohesion devices; (ii) conditional probabilities - decides the appropriateness of either using a hypothesis with a lexical cohesion or not, based on the probability of occurrence of a lexical cohesion device y, given the previous lexical cohesion

device *x*; and (iii) mutual information triggers - includes mutual information considering that *x* triggers *y* and, therefore, the chance of *y* occurring given *x* is the same as *x* triggering *y*. They integrate the information of three models as features in the decoding phase and, therefore, the documents are first translated at sentence-level and the discourse information is added later on. All models improve over the baseline (in terms of BLEU) with the mutual information triggers being the best model. Ben et al. (2013) explore a bilingual lexical cohesion trigger model that also considers lexical cohesion devices in the source side. Their results are better than the baseline in terms of BLEU.

Related Work on MT Evaluation and QE

Wong and Kit (2012) use lexical cohesion metrics for MT evaluation at document level. The authors consider the repetition of words and stems for reiteration and the use of synonyms, near-synonyms and superordinate for collocations. The integration of these metrics with the traditional BLEU, TER and METEOR are also studied. Results are compared with human assessments using Pearson's *r* correlation. The highest correlation is acquired when METEOR and the discourse features are combined.

Scarton and Specia (2014a) follow Wong and Kit (2012) and explore lexical cohesion for document-level QE of machine translated texts in English-Portuguese, English-Spanish and Spanish-English. Results show slight improvements over a baseline (Chapter 4). Finally, Gong, Zhang, and Zhou (2015) combine topic models and lexical cohesion with BLEU and METEOR for two datasets of Chinese-English machine translations. Results show that the combination of BLEU or METEOR with topic models or with lexical cohesion yield higher correlation scores with human judgements than the traditional metrics alone.

In this thesis we explore lexical cohesion following (Scarton and Specia, 2014a), considering word and stem repetitions as features for document-level QE (Chapter 4).

3.3 Coreference Resolution

Coreference occurs through coherence realisation (Hobbs, 1979). As mentioned before, coherence encompasses an overlap of information among different discourses in a given document. In order to understand coreference, we need to introduce the concept of **referring expressions**.

Referring expressions refer to objects that were previously introduced in the discourse or are introduced for further discussion (Ramsay, 2004). They can also make reference to

some world knowledge that both reader and writer share (Stede, 2011). Coreference is, then, the phenomenon where different referring expressions refer to the same object. For example Stede (2011, p. 40):

(viii) "A man named Lionel Gaedi went to the Port-au-Prince morgue in search of his brother, Josef, but was unable to find his body among the piles of corpses that had been left there."

The underlined expressions in Example (viii) are referring expressions that have coreferences. *Josef* refers to *his brother*, while *his* refers to *A man named Lionel Gaedi*. The *his* (in *his body*) refers to *Josef* and *his brother*. The referring expression *there* refers to *the Port-au-Prince morgue*.

Anaphors, pronouns, definite and indefinite noun phrases (NP) are examples of referring expressions. Resolving a coreference means to make explicit the connections among referring expressions that point to the same object (e.g. finding the NP that a pronominal anaphor refers to).

3.3.1 Anaphora Resolution

Referring expressions that point back to expressions already presented in the text are called **anaphora**. An anaphor is the referring expression that **refers to** an antecedent (another referring expression that was mentioned previously). For example Mitkov (2004, p. 267):

(ix) "**Sophia Loren** says she will always be grateful to Bono. The actress revealed that the U2 singer helped her calm down during a thunderstorm while travelling on a plane."

All the underlined expressions in Example (ix) refers to the same entity (*Sophia Loren*) forming a **coreferential chain**. There are different types of anaphora: pronominal, lexical, verb, adverb and zero. In this thesis, we explore the pronominal anaphora that are realised by personal, possessive, reflexive, demonstrative or relative pronouns. This is the most common type of anaphora and can be identified by NLP tools with more confidence than other anaphora types.

Anaphora resolution is the task in NLP that aims to identify the antecedent of a given anaphor. Although it can be simple to humans, this task poses several challenges for automatic processing. Firstly, anaphors need to be correctly identified. This means that anaphoric pronouns, NP, verbs and adverbs need to be separated from non-anaphoric terms. Secondly,

the candidates for antecedents need to be selected. Finally, each anaphor is resolved according to the candidate list.

Finding the right antecedents among all the candidates can be a very difficult for an automatic system, which makes the task of anaphora resolution very problematic. For example, in the sentence "The soldiers shot at the women and *they* fell." Mitkov (2004, p. 272), it is not easy to resolve the anaphor *they*. For humans, it can be trivial since it is expected that after getting shot *the women fell*. However, this involves world knowledge, which automatic approaches do not have access to.

A popular way of building systems for automatic anaphora resolution rely on the use of corpora annotated with coreference chains and ML approaches.[2]

Related Work on MT

LeNagard and Koehn (2010) explore anaphora resolution techniques applied to English-French SMT. The framework consists in: (i) identifying the pronouns in the source side; (ii) applying techniques of anaphora resolution to find the noun phrases that resolved each anaphoric pronoun; (iii) finding the word or noun phrase aligned to this noun phrase in the target language; (iv) identifying the gender of the word in the target language and; (v) annotating the gender in the source pronoun. Therefore, nothing is changed in the SMT system, it is only trained with annotated data. No improvements in terms of BLEU are found and the authors claim that the low performance of the anaphora resolution systems has impacted their results. Guillou (2012) follows the same framework of LeNagard and Koehn (2010), but for English-Czech. In her work, the pronouns are annotated with gender, number and animacy information. However, again, no improvements are found against a baseline system. Recently, Novák, Oele, and van Noord (2015) explore coreference resolution systems for improving a syntax-based MT system. Their approach also includes the use of rules that choose which pronoun in the target has antecendent anaphoric information. Results for English-Dutch and English-Czech show slight improvements in terms of BLEU over a baseline.

Hardmeier and Federico (2010) explore pronominal anaphora resolution for English-German translations, also using an anaphora resolution system to annotate the pronouns and their antecedents. However, the information from the anaphora resolution system is fed into the decoder in order to take the antecedent of a given pronoun into account for its translation. No significant improvement in terms of BLEU is found. However, when evaluating the precision and recall of pronoun translations, their system outperforms the baseline.

[2]More details about work on anaphora resolution can be found in Mitkov (2004) and Stede (2011).

Hardmeier (2014) also explores pronominal anaphora resolution for MT, but in a more specialised way. Firstly, a neural-network-based model to predict cross-lingual pronouns is built. This model performs pronoun resolution and pronoun labelling in one step. An important feature of this neural network is that it only uses parallel corpora without relying on tools for anaphora resolution. This information about pronouns is then included into a document-level decoder (also described in Hardmeier (2014)). This approach improves over the baseline in terms of BLEU in one of their test sets (English-French).

Machine translation of pronouns has been addressed as a translation shared task. The first was organised in 2015 (Hardmeier et al., 2015) having six participating systems. Although the systems did not outperform the baseline in terms of BLEU-style metrics, they showed better results when metrics that take into account pronoun translations were used.

In this thesis we only address pronominal anaphora by counting the number of personal pronouns in the documents and use this as features for document-level QE (Scarton and Specia, 2015) (Chapter 4).

3.3.2 Local Coherence

Entity-based Local Coherence (Barzilay and Lapata, 2005) is a model of coherence based on the **Centering Theory** (Grosz, Joshi, and Weinstein, 1995). This theory assumes that a text segment that encompasses information about a single discourse entity (coreferent Noun Phrases (NPs)) is more likely to be coherent than a segment with information about several discourse entities. It also defines the concept of *focus*: the most important discourse entity in a sentence; and *transition*: how the *focus* is passed from a sentence into another. The syntactic realisation of the *focus* is also discussed: such entities are often found as subject or object and are often referred by the use of anaphors.

In the method proposed by Barzilay and Lapata (2005) the text is represented as a matrix where the **columns contain the discourse entities** and the **rows correspond to the sentences**. The cell contains information about the syntactic realisation of the discourse entities in the sentences. In order to create the grid, the coreferent entities need to be clustered and then coreference resolution systems are applied. By using this model, the coherence of texts is assessed based on how likely the transitions that the entities have in adjacent sentences are, while represented as rows in the matrix.

Guinaudeau and Strube (2013) propose a graph-based method for local coherence by simplifying the method of Barzilay and Lapata (2005). Their method is based on a bipartite graph which avoids data sparsity and is a more robust way of coherence representation that

can efficiently model relations between non-adjacent sentences. All nouns are considered as discourse entities and the graph maps the connections among the sentences where the entities appear. Coherence is then derived from the number of shared edges in the graph. Guinaudeau and Strube (2013) achieve similar results to Barzilay and Lapata (2005).

A model for local coherence based on syntactic patterns was proposed by Louis and Nenkova (2012). This model assumes that local coherence can be defined by syntactic patterns between adjacent sentences. They also propose an approach for global coherence that assumes sentences with similar syntactic realisation have similar communicative goal.

Sim Smith, Aziz, and Specia (2016a) developed a tool (Cohere) that extracts a coherence score for documents, by using the models from Barzilay and Lapata (2005), Louis and Nenkova (2012) and Guinaudeau and Strube (2013). They also propose a new method for local coherence based on Louis and Nenkova (2012)'s model and IBM-1 alignments, that considers the alignments between adjacent sentences as a latent variable.

Related Work on MT

Sim Smith, Aziz, and Specia (2016b) explore the models implemented in Cohere (Sim Smith, Aziz, and Specia, 2016b) to evaluate the coherence of machine translated documents, considering German-English, French-English and Russian-English language pairs. Their hypothesis was that machine translated documents would show worse coherence scores than reference documents. The IBM-1 and the graph-based models show the best results since they scored machine translated documents worse than reference translations.

Related Work on QE of MT

Scarton et al. (2016) use Cohere to extract coherence features for document-level QE, building a system submission for the WMT16 QE shared task. The graph based approach of Guinaudeau and Strube (2013) was used to extract coherence scores for source and target documents. Such features were combined with our discourse features presented in Chapter 4. The best model in the test set uses coherence information from source and target.

In this thesis although we have explored coherence scores for document-level QE in Scarton et al. (2016), such approach did not show improvements over other methods implemented by us (such as the ones presented in Chapter 6). Therefore, given the poor results obtained compared to other features and the complexity of running Cohere for languages other than English, we decided to not include these features in the experiments presented in this thesis.

3.4 Small Units of Discourse

According to Stede (2011), small units of discourse are responsible for the logical component that makes a document coherent. **Coherence relations** connect text spans together inside a document, according to semantic or pragmatic factors. Such relations can be explicit or implicit. For example:

(x) "John took a train from Paris to Istanbul, **but** he never arrived." Stede (2011, p. 79)

(xi) "John took a train from Paris to Istanbul. He has family there." Stede (2011, p. 84)

Whilst in Example (x) the discourse connective *but* is explicitly connecting the two text segments (and establishing a contradictory relation between them), the connection between the two sentences in Example (xi) is not marked. We can infer a causal relation between the sentences in Example (xi), which is that *John* did something **because** he had reasons for that, although there is nothing explicitly defining this causal relation.

Discourse connectives play an important role in the task of identifying discourse relations. They explicitly define why two segments of texts are connected.

Regarding discourse relations themselves, RST (Mann and Thompson, 1987) is a linguistic theory that has been widely used to describe such relations, not only in linguistic studies but also in computational linguistics, in order to build discourse parsers (more details about RST are presented in Section 3.4.2).

A crucial concept in RST is the definition of Elementary Discourse Units (EDUs). EDUs are the text segments connected by discourse relations. The task of automatically identifying EDUs is challenging on itself, since EDUs need to be understood in order to define which coherence relation connects them. Moreover, these discourse relations can or cannot be explicitly marked (as shown in Examples (x) and (xi)).

3.4.1 Discourse connectives

Discourse connectives are composed by one (e.g. *therefore*) or more words (e.g. *even though*) that can only be disambiguated by the use of semantics and pragmatics. They belong to a closed-class of words and they are used to **connect sentences or EDUs in a logical way**. Example (xi) shows the word *but* working as a discourse connective, connecting two EDUs *John took a train from Paris to Istanbul* and *he never arrived*. The relation between the two EDUs is a contradiction: the second EDUs is contradicting the first one.

Therefore, discourse connectives can be classified by the type of the relation that they establish (*but* could be classified as a *Contradiction* connective or, in a more coarse classification as a *Comparison* connective). Exhaustive lists of connectives have been built, with connectives being classified in fine-grained or coarse classes. Nevertheless, connectives can be ambiguous and belong to different classes (or even be used in non-discourse relations). Examples (xii) and (xiii) show cases where the connective *since* is used in two different senses. In Example (xii), *since* is classified as a *Temporal* connective, while in Example (xiii), it is classified as *Contingency*.[3]

(xii) "They have not spoken to each other **since** they saw each other last fall."

(xiii) "I assumed you were not coming **since** you never replied to the invitation."

On the other hand, Examples (xiv) and (xv) show the word *and* being used as a connective and as a non-discourse marker. While in Example (xiv) *and* is classified as *Expansion*, in Example (xv) *and* does not establish any discourse relation.

(xiv) "John likes to run marathons, **and** ran 10 last year alone."

(xv) "My favorite colors are blue **and** green."

An important resource containing information about discourse connectives is the Penn Discourse Treebank (PDTB) (Prasad et al., 2008), a corpus annotated with discourse relations. In this corpus, discourse relations are annotated as both explicit and implicit. When a relation is identified, the arguments of the discourse relations (EDUs) are also annotated, following the PropBank annotation style (e.g. Arg1, Arg2, etc.) (Palmer, Gildea, and Kingsbury, 2005).

A system aiming to automatically identify and classify discourse connectives should be able to distinguish between discourse connectives and non-discourse markers and correctly classify the discourse connectives. The Discourse Connectives Tagger developed by Pitler and Nenkova (2009) is an example of such a type of system that classifies discourse connectives in *Expansion*, *Contingency*, *Comparison* or *Temporal*. Non-discourse markers are also identified. The authors used the PDTB discourse annotations in order to train a maximum entropy classifier that aims to distinguish discourse connectives from non-discourse connectives and classify discourse connectives in one of the four classes mentioned previously. As features, Pitler and Nenkova (2009) explore syntactic information along with the string of the connectives. For the task of distinguishing discourse connectives from non-discourse

[3]Examples (xii), (xiii), (xiv) and (xv) are extracted from http://www.cis.upenn.edu/~nlp/software/discourse.html.

connectives they achieve 96.26% of accuracy, whilst for the sense classification task results show 94.15% of accuracy.

Related Work on MT

Meyer et al. (2011) train a classifier for connectives on the PDTB in order to improve machine translations for English-French. The connectives are classified into six classes: *Temporal*, *Cause*, *Condition*, *Contrast*, *Concession* and *Expansion*. Ambiguous connectives are classified in multiple classes directly into the training data. In this way, the MT system learns translated rules from the annotated version of the corpus with connectives, which leads to a slight improvement in BLEU over a baseline.

Meyer and Popescu-Belis (2012) propose three different ways of dealing with connectives for English-French machine translation. The first approach changes the MT system phrase-table by adding information about manual labelled connectives into it. The second uses manually annotated connectives and train a SMT system over the annotated data. Finally, the third approach considers automatically annotated connectives to train a SMT over the annotated data (the connectives classifiers are trained on the PDTB). They evaluate the different approaches in terms of BLEU and also in terms of connectives changes (number of connectives that are correctly, incorrectly or not translated). The modified phrase-table experiments show the best results for the percentage of connectives changed correctly. In terms of BLEU, no major improvement is achieved in any of the experiments.

Meyer et al. (2012) also explore the problem of discourse connectives translation for English-French. The discourse connective classifier only targets seven ambiguous connectives, because it is expected that MT systems would fail in cases of ambiguity. They integrate these labelled connectives into a hierarchical and a phrase-based SMT systems in two ways. Similarly to Meyer and Popescu-Belis (2012), their first approach includes information about discourse connectives into the phrase table. The second approach translates the annotated data directly. For evaluation, they use BLEU and a new metric called Accuracy of Connective Translation (ACT). ACT compares the candidate translations of a connective with the human reference (retrieved from a dictionary). The phrase-based version of the discourse augmented factored translation approach shows the best results for both BLEU and ACT.

Li, Carpuat, and Nenkova (2014) study connectives in order to improve MT for Chinese-English and Arabic-English. Firstly, they examine the cases where a single sentence in the source side is manually translated into more than one sentence in the target side. Chinese-English shows many cases in which sentences were translated from one to many, because finding the sentence (or even word) boundaries in Chinese is not trivial. Besides that, there is

a high correlation between bad HTER scores and longer Chinese sentences that should be translated into many English sentences. Afterwards, they evaluate explicit connectives by using an automatic discourse connectives tagger. Considering the number of connectives, only Chinese-English shows correlation with HTER (the presence of connectives was related to higher HTER). Considering the ambiguity of connectives (connectives that can have more than one sense), again only Chinese-English shows correlations with HTER (the presence of ambiguous connectives yielded to bad HTER). Finally, the authors compare the presence of discourse connectives in the translations and post-editions. In this case, they consider the following scenarios: (i) no discourse connectives in both; (ii) the same discourse connectives appear in both, in the same sense and; (iii) there is a discourse connective only in translation or only in the post-edited version. In both languages, scenario (iii) shows higher HTER than the other two.

Finally, Steele and Specia (2016) propose a method for improving the translation of implicit discourse elements in Chinese-English machine translation. Such discourse elements include discourse connectives, pronouns, and elements that often relate to pronouns (e.g. "it's"). Their approach uses manual and automatic word alignments over the training corpus to make explicit some implicit constructions in source. In summary, they annotate the source with the missing information that is presented in the target. SMT systems built with the annotated data show improvements over the baselines (in terms of BLEU).

There is no previous work that uses discourse connectives information for QE.

In this thesis we use the Pitler and Nenkova (2009)'s discourse connectives tagger in order to extract discourse connectives features for document-level QE (Scarton and Specia, 2015) (Chapter 4).

3.4.2 RST

Rhetorical Structure Theory (RST) is a theory that aims to define discourse relations among EDUs. RST is not sensitive to text size, and is flexible to identify explicit and implicit discourse relations. Relations in RST can be organised in a tree (mostly binary, although relations such as JOINT allow more than two EDUs). The arguments in a relation are classified as *Nucleus* or *Satellite*, according to their function in the text.

The graph in Figure 3.1 shows the sentence "Brown has coveted the office of Prime Minister since May 12, 1994, the fateful day when John Smith, the Labour leader in opposition, died of a heart attack." in a RST tree (automatically generated), where the ***Nucleus*** is *"Brown has coveted the office of Prime Minister since May 12, 1994, the fateful day"*, the

Satellite is *"when John Smith, the Labour leader in opposition, died of a heart attack."* and the relation between them is ELABORATION.

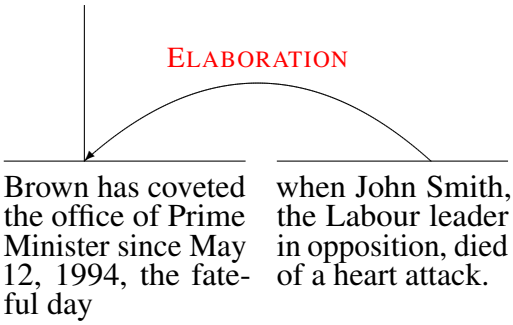

Fig. 3.1 Example of RST relation between two EDUs.

Given its feature of tree organisation, RST has been used by different purposes in NLP. Joty et al. (2013) developed an RST parser that extracts EDUs and provides the RST trees at sentence and document levels. The first step in this parser is to segment sentences into EDUs that later will be connected by the discourse relations. The EDUs are identified by using the SPADE tool (Soricut and Marcu, 2003) and the Charniak's syntactic parser (Charniak, 2000).

The second step consists in establishing inter-sentential and intra-sentential relations. Firstly, RST subtrees are built at sentence level in order to identify the intra-sentential relations. Then, the parser builds the inter-sentential relations by connecting the subtrees for each sentence into a single RST tree. Both approaches (inter and intra) are composed by two modules: parsing model and parsing algorithm. Inter-sentential and intra-sentential cases have different parsing models, whilst they share the same parsing algorithm. The parsing models are responsible for assigning a probability to every candidate RST tree and are implemented using Conditional Random Fields (CRFs). On the other hand, the parsing algorithm chooses the RST tree most likely to fit a given scenario and is implemented as a probabilistic CKY-like bottom-up algorithm. Joty et al. (2013)'s parser significantly outperforms previous work in terms of f-score.

Related Work on MT

Marcu, Carlson, and Watanabe (2000) use RST to evaluate the feasibility of building a discourse-structure transfer model towards a Discourse-Based MT system. The framework proposed in their work is divided in three parts: (i) discourse parser - to apply in the source texts; (ii) discourse-structure transfer model - to rewrite the structures in the source texts

to reflect the structures of the target and (iii) statistical module - to map source and target using translation and language models with discourse information. The authors experiment with 40 Japanese-English texts manually annotated with RST relations. They show that RST discourse trees in Japanese are considerably different from discourse trees in English, which justifies the effort of building a discourse-based MT system. For the discourse-structure transfer model, they consider several operations in order to transform the source tree into the target tree. Then, they train a classifier to learn the operations and the order of these operations needed to transform Japanese structures into English structures. After that, they conduct an evaluation step with the structures in Japanese and the transformed structures in English. They found improvements in discourse across the whole documents and across sentences. However, the identification of paragraphs boundaries was a shortcoming and they could not improve the results at paragraph level. Moreover, their model is cumbersome and time-consuming for training and tuning the system. Tu, Zhou, and Zong (2013) also explore RST trees for MT purposes for Chinese-English. They use the RST trees from parallel data to extract RST translation rules and use them in conjunction with the phrase table rules. The use of RST rules lead to improvements over a baseline, in terms of BLEU.

Related Work on MT evaluation

Guzmán et al. (2014) explore RST trees (acquired automatically by using a discourse parser) for automatically evaluating translations into English. Trees of machine translated text are compared to those in the human references by using convolution Tree Kernels, which compute the number of common subtrees between two given trees. Two different representation of the trees are used: non-lexicalised and lexicalised. Their results are compared against BLEU, TER, METEOR, NIST, Recall-Oriented Understudy for Gisting Evaluation (ROUGE), all metrics submitted to the WMT12 metrics shared task and some metrics from Asiya toolkit (Giménez and Màrquez, 2010). The lexicalised version outperforms several metrics from WMT12, BLEU, TER, METEOR and NIST at system level. When combined to other automatic metrics, the lexicalised version improves over 13 out of 19 metrics at system level. Joty et al. (2014) present a set of metrics called DiscoTK, also based on RST trees (following Guzmán et al. (2014)), although this set of metrics is tuned on human judgements and combined to Asiya metrics. DiscoTK was the best metric in the WMT14 metrics shared task[4] for four out six language pairs at segment level and for two out five language pairs at system level (Macháček and Bojar, 2014).

[4]http://www.statmt.org/wmt14/

Similar to discourse connectives, there is no previous work that uses RST information for QE.

In this thesis we use Joty et al. (2013)'s RST parser and extract different features for document-level QE (Scarton and Specia, 2015) (Chapter 4).

3.5 Discussion

In this chapter, we discussed the definitions on discourse that are important for our thesis along with a discussion about related work on MT, MT evaluation and QE using discourse. We followed Stede (2011) in terms of discourse definitions and the organisation of discourse processing in three levels: large units of discourse and topics, coreference resolution and small units of discourse (Section 3.1). Such structure allowed us to keep the definitions coherent and to keep consistency along the entire thesis (the features presented in Chapter 4 are also organised based on this structure).

The related work on MT, MT evaluation and QE was also presented following Stede's classification. Regarding large units of discourse and topics, a considerable amount of work has been done in both topic modelling and lexical cohesion aiming to improve MT (Section 3.2). However, it is still unclear how to use such information for improving MT systems without losing in performance. Moreover, in terms of evaluation, very little has been done.

In the discussion about coreference resolution (Section 3.3), we found that the majority of work done focused on anaphora resolution. In fact, almost all work done with the Docent decoder aimed to improve the translation of anaphora. Nevertheless, the impact of such approaches is still not completely understood and further evaluation (perhaps beyond BLEU) still needs to be performed. In terms of MT evaluation, there is a lack of studies, which is an open ground for future investigations. Much less work has been done regarding local coherence.

As we showed in the discussion about small units of discourse (Section 3.4), a considerable amount of work has been done in discourse connectives. However, such approaches have the same problems as the ones that address anaphora: the impact they have on MT quality it is still not understood. The use of RST was also explored by work in MT and MT evaluation, although it has several practical issues, such as the need for manually annotated data or RST parsers.

We can conclude that, although there are already initiatives that include discourse information in MT and MT evaluation, there are many open issues that still need to be explored.

With recent advances in NLP, more resources and tools to process discourse are being developed and, therefore, more work addressing this topic in the context of MT is expected to be done. Moreover, with initiatives such as Docent, the inclusion of discourse information into SMT systems is more reliable, because changing the decoder may not cover the same type of context that Docent can handle. Finally, it is important to mention that the evaluation of such discourse-aware approaches need to go beyond traditional automatic evaluation metrics in order to make a fair evaluation. In fact, related work on the use of discourse information for MT presented within this chapter show none or marginal improvements over baseline systems according to BLEU. This does not necessarily mean that such new approaches are not better than the baseline, but that the inadequate evaluation may be being performed. BLEU (and the majority of the automatic evaluation metrics) does not account for discourse and changes on discourse units (sometimes sutil) will not be taken into account. Therefore, a more reliable evaluation framework needs to be developed in order to evaluate discourse-aware MT approaches.

In the next chapter we present our work on feature engineering for document-level QE. We propose an adaptation of the sentence-level features (presented in Chapter 2) for document-level QE. We also describe how we explore the discourse information presented in this chapter as features for document-level QE.

Chapter 4

Document-level QE: Feature Engineering

As mentioned in Chapter 1, feature engineering is one of the main challenges in QE.[1] Finding the best way of using the information from source and target documents, together with external resources, is one of the bases for building a well performing QE system. In this chapter we present our study on feature engineering for document-level QE.

QE features can be classified in four classes (Specia et al., 2011). **Complexity features** use information from the source texts in order to capture the difficulty in translating it. Such features can use shallow information (e.g. number of tokens in the source text) or sophisticated linguistic information (e.g. probability of a syntactic parse tree being generated for the source text). Conversely, **fluency features** use only target information and assess how fluent a given target text is. Fluency features can also use shallow (e.g. number of tokens in the target text) or deep linguistic information (e.g. probability of a syntactic parse tree being generated for the target text). Therefore, complexity and fluency features can be extracted using similar approaches. **Confidence features** use information from the MT system used to produce the translations (usually a SMT system). This kind of features attempt to capture how confident a system is in producing a given translation. Finally, **adequacy features** use information from both source and target texts. The aim of these features is to assess whether or not the translation preserves the meaning and structure of the source. Adequacy features include simple ratios between source and target counts (e.g. ratio of number of tokens

[1] Parts of this chapter were previously published in peer-reviewed journal and conferences: Scarton and Specia (2015), Scarton (2015), Scarton, Tan, and Specia (2015), Scarton and Specia (2014a) and Scarton and Specia (2014b).

between source and target) and complex information from alignments of named entities, syntactic structures, among others.

In this thesis, we focus on complexity, fluency and adequacy features. Since we consider the translation process as a black-box and we mix different types of translation systems, including commercial and non-commercial versions, we do not have access to information about the systems used. Our adequacy features are shallow ratios between source and target, since there are no resources available in various source and target languages for more sophisticated features. Finally, the majority of our features are complexity features, given that deep linguistic features are only applied for English as the source language because of the lack of reliable resources and tools for languages other than English and the difficulty of running such resources/tools on target texts. Target texts contain many ungrammatical structures that cannot be reliably processed by traditional NLP tools, generating errors or noisy annotations. Manually correcting such errors is unfeasible and it would result in artificial data.

Our features can be divided in document-wide features and discourse-aware features. Document-aware features can be simple counts at document level or aggregation of sentence-level information. Discourse-aware features use some kind of discourse information.

Moreover, we also experiment with pseudo-reference features. Pseudo-references are translations produced by one or more external MT systems, which are different from the one producing the translations we want to predict the quality for. Such systems are used as references against which the output of the MT system of interest can be compared using string similarity metrics, such as the automatic evaluation metrics for MT. For QE, such features have been explored by Soricut and Echihabi (2010) and Scarton and Specia (2014a). In both cases, pseudo-reference features lead to better QE models.

Ideally, the quality of the MT system being used to produce pseudo-references needs to be known (and this system should usually be better than the MT system under evaluation). However, in most cases, such information is not available. In Scarton and Specia (2014b), we propose a consensus approach, where several MT systems are considered as pseudo-references. The hypothesis is that if several MT systems share the same information, this information is likely to be correct. Therefore, knowing the quality of the MT systems is not a requirement. The approach employed in our pseudo-reference features follows this consensus approach.

Finally, we also explore word embeddings as features for our QE models. The assumption is that word embeddings are able to capture cohesion information, derived from the semantic

information encoded by word embeddings. By aggregating the word vectors into a document vector we expect to extract information about word choices and lexical consistency.

In Section 4.1, we present document-wide features, which include features with shallow linguistic information such as type/token ratio, number of sentences, among others. The majority of such features are adaptations of sentence-level features that have been proven useful for QE at sentence level. All features of this kind are implemented in the QUEST++ framework. Section 4.2 presents the discourse-aware features used in this thesis. Such features contains discourse information from all different levels of discourse processing discussed in Chapter 3 and are the main contribution of this thesis in terms of features. Section 4.3 presents the word embeddings features for document-level QE. Section 4.4 presents pseudo-reference and consensus-based features. In Section 4.5, we present an analysis of our discourse-aware features, correlating them with HTER. A comparison with baseline shallow features is also described.

4.1 Document-aware Features

Document-aware features cover document information that does not involve discourse processing. Such information can be simple counts (e.g. number of sentences, number of tokens) or complex calculations at sentence level which are aggregated at document level (e.g. LM perplexity, probabilities from a syntactic parser). In this thesis we use a large set of document-aware features that we implemented and made available in the QUEST++ toolkit. These features are adaptations of sentence-level features to document-level QE.

The widely used 17 baseline features for sentence-level QE were also adapted for document level. Such features were used in both editions of the WMT QE (2015 and 2016) shared tasks in document-level QE in order to build baseline QE models (Bojar et al., 2015, 2016b). The 17 baseline features are:

1. number of tokens in the target document;
2. LM probability of target document;
3. type/token ratio (number of occurrences of the target word within the target hypothesis, averaged for all words in the hypothesis);
4. number of punctuation marks in target document;
5. number of tokens in the source document;

6. average source token length;
7. LM probability of source document;
8. average number of translations per source word in the document (threshold: prob >0.2);
9. average number of translations per source word in the document (threshold: prob >0.01) weighted by the inverse frequency of each word in the source corpus;
10. percentage of unigrams in quartile 1 of frequency in a corpus of the source language;
11. percentage of unigrams in quartile 4 of frequency in a corpus of the source language;
12. percentage of bigrams in quartile 1 of frequency in a corpus of the source language;
13. percentage of bigrams in quartile 4 of frequency in a corpus of the source language;
14. percentage of trigrams in quartile 1 of frequency in a corpus of the source language;
15. percentage of trigrams in quartile 4 of frequency in a corpus of the source language;
16. percentage of unigrams in the source document seen in a corpus (SMT training corpus);
17. number of punctuation marks in source document.

These baseline features are aggregated by summing or averaging their values for the entire document. Features number 1, 4, 5 and 17 are summed, which is the same of extracting them directly for the entire document. All other features were averaged. Some features, such as Feature 1, could be directly extracted for the entire document, without needing to make sentence splits. However, since such features were already implemented into QUEST++ for sentence level, we made use of the existing implementation in order to keep the code modular. On the other hand, features like numbers 2 and 7 could not be implemented for the entire document. In the specific case of LM features, the values would be so small that would led to underflow. Therefore, we kept the implementation of LM features at sentence level and averaged the values for all sentences within the document.

Although averaging over sentence level scores is a simplistic approach, mainly because such an averaged score could be biased by outliers, in practice the majority of the baseline features extracted for our data do not show outliers. The only averaged features more sensitive to outliers are 2, 7, 8 and 9, which are features that could not be extracted for the entire document for the reasons given in the previous paragraph.

The majority of the baseline features are complexity features. Features numbers 1, 2, 3 and 4 are the only fluency features. Although there are no adequacy features in the baseline, the other features available in QUEST++ contains several adequacy features (e.g. absolute difference between number tokens in source and target, normalised by source length).

An exhaustive list of document-aware features with their descriptions is given in Appendix A. Examples of these features are:

- Complexity features:
  - average number of translations per source word in the document (threshold: prob >0.01/0.05/0.1/0.2/0.5);
  - average word frequency: on average, the number of times each type (unigram) in a source document appears *n* times in the corpus (divided by frequency quartiles);
  - LM perplexity of the source document (with and without end of sentence marker).
- Fluency features:
  - percentage of content words in the source document;
  - average of PCFG parse log likelihood of source sentence;
  - LM perplexity of the target document (with and without end of sentence marker).
- Adequacy features:
  - absolute difference between the number of tokens in the source document and the number of tokens in the target document, normalised by source document length;
  - ratio of the number of nouns in the source document and the number of nouns in the target document.

4.2 Discourse-aware Features

Discourse-aware features use discourse information from source and/or target documents for the task of QE. The intuition behind the use of such features comes from the fact that discourse is a phenomena that happens beyond sentence level. Therefore, document-level QE models should also explore discourse information. In addition, traditional MT systems fail in providing translations that take into account document-aware information, since documents are translated sentence-by-sentence and, therefore, we expect issues of this type in the MT

systems outputs. The main challenge in extracting discourse-aware features is to find reliable sources of discourse information. Although human annotated data would probably be the best resource, it is not available for a large amount of data and it is domain dependant.

As it is done for other levels of QE, the way to overcome the lack of discourse-annotated data that could be used for document-level QE was to use automatic tools. As mentioned in Chapter 3, discourse parsers and taggers have been made available over the years. As a consequence, the features that we present in this section make use of different tools in order to extract discourse-aware information. We present the features that we implemented by grouping them considering the classification proposed by Stede (2011) which we described in Chapter 3. It is worth noting that the majority of the discourse-aware features are complexity features (with the exception of lexical cohesion, LSA and LDA). The reason for this is that the resources and tools used are available only for English and they assume correct documents as input. Therefore, such features are extracted only when English is used as the source language.

4.2.1 Large Units of Discourse and Topics

As discussed in Chapter 3, lexical cohesion and topic modelling are aspects of large units of discourse and our features cover both.

Lexical Cohesion

As defined in Chapter 3, Lexical Cohesion (LC) is related to word repetition (reiteration) and collocation. Regarding the evaluation of documents, this aspect of discourse was explored as features for Readability Assessment (Graesser et al., 2004) and reference-based MT evaluation (Wong and Kit, 2012). Following this work, we propose the first set of features for QE using LC.

These features are based in word repetition only, since repetitions are language independent and can be used for both source and target texts. Synonyms and other types of semantic relations require resources like WordNet (Fellbaum, 1998) that are not available for most languages. Besides that, the coverage of these kinds of resource vary across languages, and it could influence the reliability of the features.

Therefore, our features are counts of word, lemma or noun repetitions across a document:

- **Average word repetition:** for each content word, we count its frequency in all sentences of the document. Then, we sum the repetition counts and divide it by the total number of content words in the document (Equation 4.1, where w_i is the i^{th} word in a

document, N is the total number of content words in a document and $freq$ is a function that outputs the frequency of w_i in a document). This is computed for the source and target documents, resulting in two features: one for source and another for target.

$$\frac{\sum_{i=1}^{N} freq(w_i)}{N} \tag{4.1}$$

- **Average lemma repetition:** the same as above, but the words are first lemmatised (two features).
- **Average noun repetition:** the same as word repetition, but only nouns are taken (two features).
- **Ratio:** ratio of source and target word, lemma or noun repetition (three features).

A document that presents high scores for repetition is expected to have high lexical cohesion (at least in terms of lexical cohesion). The features presented above are implemented into the QUEST++ toolkit. The lemma and part-of-speech (POS) tags (needed to identify content words) were extracted by using the TreeTagger[2] tool (Schmid, 1994).

Topic Modelling

For discourse information achieved via topic modelling, we propose two sets of features. The first one accounts to the internal cohesion of the documents by building a LSA matrix for each document and considering the matrix of words $\times$ sentences. Features are extracted by measuring the similarity between the sentences (columns of the matrix). The second set of features is based on the work of polylingual LDA for QE at sentence level (Rubino et al., 2013). The features are divergence scores between the polylingual distribution of source and target documents.

LSA cohesion The LSA method (presented in Chapter 3) aims to capture the topic of texts based on the words that these texts contain. It is a robust method where texts can be full documents, paragraphs or sentences. The matrix X (Equation 3.1) is, then, built with $words \times documents$, $words \times paragraphs$, $words \times sentences$, etc. In the case of $words \times sentence$ (which we use in our experiments), each cell contains the frequency of a given word in a given sentence.

[2] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

LSA was originally designed to be used with large corpora of multiple documents (topic modelling). In our case, since we were interested in measuring coherence within documents, we computed LSA for each individual document through a matrix of $words \times sentences$ in the document.

We computed LSA using a package for Python,[3] which takes word stems and sentences to build the matrix X. We modified a few things to make it more appropriate for our purposes. The main modification refers to the TF-IDF transformation. This transformation is made to smooth the values of the high frequency words, in order to keep all words normalised according to their frequency in a given corpus. However, in our case the salience of words in sentences was important and therefore this transformation was removed. The following features for QE are extracted as follows:

- **LSA cohesion - adjacent sentences - Spearman correlation**: for each sentence in a document, we compute the Spearman's ρ correlation coefficient between its word vector and the word vectors of its adjacent neighbours (sentences which appear immediately before and after the given sentence). For sentences with two neighbours (most cases), we average the correlation values. After that, we average the values for all sentences in the document in order to have a single figure for the entire document.
- **LSA cohesion - adjacent sentences - Cosine distance**: the same as above, but applying cosine distance instead of Spearman's ρ correlation.
- **LSA cohesion - all sentences - Spearman correlation**: for each sentence in a document, we calculate the Spearman's ρ correlation coefficient of the word vectors between a given sentence and all the others. Again we average the values for all sentences in the document.
- **LSA cohesion - all sentences - Cosine distance**: the same as above, but applying cosine distance instead of Spearman's ρ correlation.

Higher correlation scores are expected to correspond to higher text cohesion, since the correlation among word vectors of sentences in a document is related to how closely related the words appear in the document (Graesser et al., 2004). Example (xvi), therefore should present higher LSA scores than Example (xvii), because the sentences in the former are related to a single topic (a "pleasant day"), whilst the sentences in the latter do not refer to the same topic.[4]

[3] https://github.com/josephwilk/semanticpy

[4] Examples are extracted from Coh-Metrix Documentation: http://cohmetrix.com/.

(xvi) "The field was full of lush, green grass. The horses grazed peacefully. The young children played with kites. The women occasionally looked up, but only occasionally. A warm summer breeze blew and everyone, for once, was almost happy."

(xvii) 'The field was full of lush, green grass. An elephant is a large animal. No-one appreciates being lied to. What are we going to have for dinner tonight?"

As opposed to lexical cohesion features, LSA features are able to find correlations among different words, which are not repetitions and may not even be synonyms, but are still related (as given by co-occurrence patterns).

Polylingual LDA We adapt the work of Rubino et al. (2013) on polylingual LDA for sentence-level QE in order to use it for document-level QE. In Rubino et al. (2013), the polylingual method consists in extracting aligned monolingual topic models. Therefore, source and target sentences are processed separately and are aligned afterwards.

In the case of document-level QE, instead of building a topic model for each sentence, we extract a topic model for each document. The topic models are extracted using MALLET toolkit[5] and the LDA features are computed inside QUEST++ toolkit. An intermediate script (developed by Rubino et al. (2013)) is used to make the topic models readable by QUEST++. In QUEST++, two features are computed:

- Kullback-Leibler (KL) divergence between a source document and a target document topic distribution;
- Jensen-Shannon (JS) divergence between a source document and a target document topic distribution.

Both KL and JS divergence scores are based on probabilistic uncertainty and do not require the probability distributions to be represented in Euclidean space (Rubino et al., 2013). Considering a vector s_i representing the topic model probability distribution of a source document, a vector t_i representing the topic model probability distribution of a target document and n being the number of dimensions of the topics,[6] the KL divergence score is defined by Equation 4.2.

$$KL(s,t) = \sum_{i=1}^{n} s_i \log \frac{s_i}{t_i} \tag{4.2}$$

[5]http://mallet.cs.umass.edu

[6]By default, QUEST++ considers 50-dimension vectors

JS is a symmetric version of KL, calculated according to Equations 4.3.

$$JS(s,t) = \frac{1}{2}(\sum_{i=1}^{n} s_i \log \frac{2s_i}{s_i+t_i} + \sum_{i=1}^{n} t_i \log \frac{2t_i}{s_i+t_i}) \tag{4.3}$$

The features extracted from the polylingual LDA are expected to be a good indicator of adequacy at document level. They are able to capture whether or not source and target aligned documents are similar in terms of the topics they address.

4.2.2 Coreference Resolution

Regarding coreference resolution (explained in Chapter 3), two main discourse processing approaches are employed: anaphora resolution and local coherence models. Our features attempt to count anaphora only.[7]

Anaphora

We count the number of personal pronouns in a document and use this as a feature for QE. Pronominal anaphora are expected to be a problem for MT systems since they perform translation disregarding document-aware information. If pronouns like *they*, *he* or *it*, for example, appear in a document, one expects that there will be an antecedent that resolves them, which may or may not be done by the MT system. Moreover, when translating into morphologically richer languages, pronouns like *they* and *it* should have an explicit gender (the same gender as the antecedent).

Therefore, the presence of pronouns can make the document more complex in MT. This feature is thus related to text complexity and is expected to work better when applied in the source side. It is worth mentioning that a more reliable approach would be to compare source and target pronouns (e.g. calculating the ratio between the number of pronouns in source and target documents). However, there is a lack of reliable tools for languages other than English.

In our implementation we used the output of Charniak's parser for pronoun identification. Such parser was also used in the pre-processing step of the features presented in Section 4.2.3 and, therefore, its results were also used for our pronoun feature. Pronouns were identified by looking at the tag "PRP" of the Charniak's parser .[8]

[7] As explained in Chapter 3, local coherence models were studied in a joint work with Karin Sim Smith, although no improvement were found. Please refer to Scarton et al. (2016) for results.

[8] Since Charniak's parser is only available for the English language, this feature can only be extracted for English.

4.2.3 Small Units of Discourse

Discourse connectives, EDUs and RST relations are classified as small units of discourse. We propose several features that account for the three phenomena.

Discourse Connectives

Discourse connectives can be seen as the "glue" that ties sentences together in a logical way (as presented in Chapter 3). The simple presence of connectives can be already seen as a sign of coherence in a document. Our features count the number of discourse connectives in a document.

Connectives are identified by using the Discourse Connectives Tagger from Pitler and Nenkova (2009).[9,10] This *tagger* tags connectives in one of the four classes: *Expansion*, *Contingency*, *Comparison* and *Temporal*. Non-discourse connectives are also tagged as *non-discourse*. The features extracted by this information are:

- Total number of connectives (including non-discourse connectives);
- Number of *Expansion* connectives;
- Number of *Contingency* connectives;
- Number of *Comparison* connectives;
- Number of *Temporal* connectives;
- Number of *Non-discourse* connectives.

As with the pronominal anaphora feature, these features can be seen as complexity features. Discourse connectives can be ambiguous (such as "then") which can make the translation process harder.

EDUs

As presented in Chapter 3, EDUs are the minimum units of text that assume some discourse role. For example, in RST, the discourse annotations are made at the EDUs level. In the following example, we can see several clauses breaks marked by *EDU_BREAK*:

[9]The Charniak's syntactic parser is used to pre-process the data.
[10]In our implementation, these features can only be extracted for English language.

(xviii) "However , **EDU_BREAK** despite the economic success , **EDU_BREAK** it has become increasingly clear **EDU_BREAK** that Blair was over ."

EDU breaks are marked using the Discourse Segmenter module of the Discourse Parser developed by Joty et al. (2013). This module uses the outputs of the Charniak's parser and the EDUBREAK module of the SPADE tool (Soricut and Marcu, 2003) in order to break sentences into EDUs. Our feature counts the number of breaks in the entire document and it is a complexity feature. Documents with a high number of EDU breaks may be composed by complex structures that are harder to translate.[11]

RST

As explained in Chapter 3, RST is theory that associates small units (EDUs) according to discourse relations. Each small unit receives a discourse role and the document (or sentence) is represented as a discourse tree. An example of a RST ELABORATION relation is presented in Chapter 3 (Figure 3.1). Our features using RST information are:

- number of *Nucleus* relations;
- number of *Satellite* relations;
- height of the RST tree;
- number of subtrees in the RST tree.

RST relations are also extracted by using Joty's parser.[12] This parser is able to annotate RST trees at sentence and document levels. At document level, the trees go from the smallest units (EDUs) to sentences and paragraphs, until they reach the full document. At sentence level, the trees model intra-sentence discourse relations.

Our features use the document-level tree from Joty's parser and are complexity features. Our assumption is that a higher number of *Nucleus* and *Satellite* relations refer to the use of too many discourse units and, therefore, leads to a complex document. Moreover, if the number of subtrees and the height of the tree are also high, they can also be a sign of a complex document.

[11] Given the language requirements of the tools used, our feature can only be used for English.
[12] Given the requirements of Joty's parser, these features are only implemented for the English language.

4.3 Word Embeddings Features

As was mentioned in Chapter 3, word embeddings can be viewed as a topic modelling approach.However, due to the special characteristics of these features and the lack of consensus on whether they are compatible to topic modelling approaches or not, we separate them from the discourse-aware features. For document level we aggregate all word embeddings into a single vector. Such an aggregated vector (that can be the sum or average of all word vectors in a document, for example) is a representation of all words in the document and can encode information such as lexical cohesion and topics. We used the approach presented in Scarton et al. (2016) where word embeddings are averaged. We experimented with other *pooling* approaches such as sum, minimum and maximum, however, averaging the vectors showed the best results.[13] We train CBOW models for different languages using the *word2vec* tool.[14] The models learn word embeddings with 500 dimensions.

It is worth mentioning that we also experimented with the *doc2vec* approach (Le and Mikolov, 2014), using the implementation available from the `gensim` framework[15] (Řehůřek and Sojka, 2010). However, experiments with the document vectors did not show improvements over the approach of averaging word vectors.[16]

4.4 Consensus Features

As mentioned previously, pseudo-reference features showed promising results in previous work in document-level QE. However, this kind of features can not always be applied. In order to use pseudo-references, translations by MT systems other than the system we want to predict quality for need to be available. In this scenario, traditional evaluation metrics can be applied. Moreover, according to Soricut and Echihabi (2010), the quality of the MT systems used as pseudo-references needs to be known. Then, the comparison of the output of the MT system under investigation against the pseudo-references would be guided by the quality of the off-the-shelf systems (e.g. if a pseudo-reference system is known to be good we can infer that a system under investigation is also good if the output of both systems are close).

[13] Part of this work was done in conjunction with Dr Kashif Shah, who implemented the script for aggregating the word embeddings.

[14] https://code.google.com/archive/p/word2vec/

[15] https://radimrehurek.com/gensim/models/doc2vec.html

[16] Although Le and Mikolov (2014) presents *doc2vec* as the best approach for extracting document vectors, it is now known that the results were not reproducible due to a problem in the data preparation (T. Mikolov, personal communication (email), March, 2016). More about this topic can be found at https://groups.google.com/forum/#!topic/word2vec-toolkit/Q49FIrNOQRo.

In Scarton and Specia (2014b), we propose an approach where several MT systems are considered and BLEU, TER, METEOR and two extra syntactic metrics from Asiya toolkit[17] Giménez and Màrquez (2010) are used as pseudo-reference metrics. The authors assumption is that if several MT systems show the same translation, there is a higher chance that this translation is right. Therefore, the quality of the MT systems does not need to be known. The automatic metrics were calculated at sentence level and used as features for sentence-level QE. Results in the 2014 WMT QE shared task were promising.

In this thesis we present results by using pseudo-references and in a consensus-based approach as used by Scarton and Specia (2014b). It is worth mentioning that this approach was only applied when translations from different MT systems were available. Consensus features can be viewed as a kind of fluency features, given that they compare the target with different MT outputs in the target language, aiming to find a consensus between them.

4.5 Feature Analysis

In our work published in Scarton and Specia (2015), we performed an analysis of a representative set of discourse-aware features by correlating them with HTER and comparing such results with the correlation of the 17 baseline features for document-level QE and HTER. Our purpose was to find whether or not discourse-aware features would be useful for predicting document-level quality. We report these experiments in this thesis since they show an useful analysis of our discourse features. In this study we also explored discourse features for English as the target language. We did not keep this kind of feature because of the problems encountered in pre-processing the data with NLP tools. Some translated structures needed to be manually changed in order to apply the tools, which proved to be very time-consuming. Correcting translated structures is also not ideal for the QE task, since the evaluation of MT system outputs should not require humans correcting the translation. Therefore, we did not use discourse features for the target language in the experiments presented in Chapters 5 and 6.

We used two corpora with machine translations and post-editions: the LIG (Potet et al., 2012) and Trace corpora (Wisniewski et al., 2013). LIG contains 10,881 French-English (FR-EN) machine-translated sentences (and their post-editions) from several editions of WMT translation shared tasks (news documents). The document boundaries were recovered and the HTER was calculated at document level.[18] 119 documents were analysed. Trace

[17] http://nlp.lsi.upc.edu/asiya/

[18] We are thankful to Karin Smith for generating the document mark-ups.

contains 6,693 FR-EN and 6,924 English-French (EN-FR) machine-translated sentences with their post-editions. We used 38 documents recovered from the WMT and Technology, Entertainment and Design (TED) Talks EN-FR sets.[19] The HTER values were also calculated at document level. Only the phrase-based SMT outputs were considered.

We use a subset of the discourse-aware features we implemented, that covers all the discourse levels:[20]

- Large units of discourse and topics (Section 4.2.1):
  - **LC:** lexical cohesion features at document level.
  - **LSA Cohesion:** LSA cohesion features at document level.
- Coreference resolution (Section 4.2.2):
  - **Pronouns:** counts of pronouns.
- Small units of discourse (Section 4.2.3):
  - **Connectives:** counts of connectives.
  - **Discourse unit segmentation (EDU break):** number of breaks (EDU). An example of a sentence broken into EDUs is the following:
  - **RST relations:** number of Nucleus and number of Satellite relations.

For comparison, the 17 baseline features from QUEST++ are also included in our experiments (see Section 4.1 for the list of baseline features). The analysis is done on the English side only for all features (given that some discourse features can only be applied for this language). Therefore, for the LIG corpus we only consider the baseline features that are extracted for the target document and, for the Trace corpus, the baseline features that are extracted from the source. Spearman's ρ and Pearson's r correlation coefficients are used for our correlation analysis (a p-value smaller than 0.05 is considered significant for both correlation scores). Spearman's ρ measures the monotonic relationship between the features and HTER, whilst Pearson's r assesses the linear relationship between the features and HTER. Therefore, while Spearman's ρ is more adequate for ordinal data and less sensitive to outliers, Pearson's r assumes that there is a linear relationship between the two variables compared.

[19] The other sets did not have document-level mark-ups.

[20] This subset was selected based on the type of information that the discourse features encode. Similar features that showed similar results were not present on this analysis.

To better evaluate whether discourse features correlate with HTER scores, besides applying the analysis to the entire corpus, the LIG corpus was divided into four bins.[21] The bins show how the features behave in different quality intervals according to HTER. The documents are sorted according to HTER and split into bins as follows:

- 10 documents: five documents with the best five HTER scores and the five documents with the worst five HTER scores.
- 20 documents: ten documents with the best ten HTER scores and the ten documents with the worst ten HTER scores.
- 40 documents: 20 documents with the best 20 HTER scores and the 20 documents with the worst 20 HTER scores.
- 80 documents: 40 documents with the best 40 HTER scores and the 40 documents with the worst 40 HTER scores.

Figure 4.1 and Figure 4.2 show the correlation scores in terms of Pearson's r and Spearman's ρ, respectively, for the LIG corpus. Results for both metrics are considerably similar (with rare exceptions). Since the analysis was done for the target side only, the QUEST++ features used were *QuEst* 1 to *QuEst* 4 (the numbers of QUEST++ features follow the baseline list presented in Section 4.1) . For the bin with 10 documents, the discourse features *RST - Nucleus*, *RST - Satellite* and *EDUs*, together with the *QuEst* 1 feature, show the best correlation scores according to both Pearson's r and Spearman's ρ. For the bin of 20 documents, *QuEst* 2 and *RST - Nucleus* show the highest Pearson's r correlation scores with HTER (above 0.37). The highest Spearman's ρ correlation score is shown by *Pronouns*.

For bins with 40, 80 and all documents (119), the *LC - Argument Target* feature shows the highest Pearson's r and Spearman's ρ correlation scores (around -0.354, -0.23 and -0.20 respectively). Note that, in this case, the correlation scores are negative, but they still indicate correlation between quality and feature. In fact, a negative correlation is expected because higher values for the *LC - Argument Target* feature mean higher document cohesion and thus lower HTER scores.

As expected, both Pearson's r and Spearman's ρ correlation scores are higher when moving from all documents to the 10-document bin. However, this was not the case for all features. In fact, it is possible to observe in Figures 4.1 and 4.2 that even taking the extreme quality values only leads to larger correlations for some discourse features. In the

[21] The portion of the Trace corpus used here was too small to be split into bins (only 38 documents).

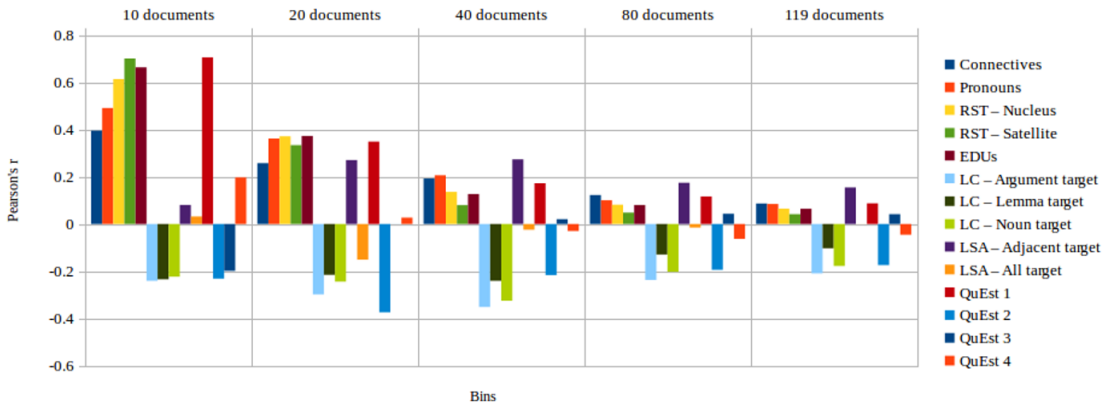

Fig. 4.1 Pearson's *r* correlation between target features and HTER values on the LIG corpus.

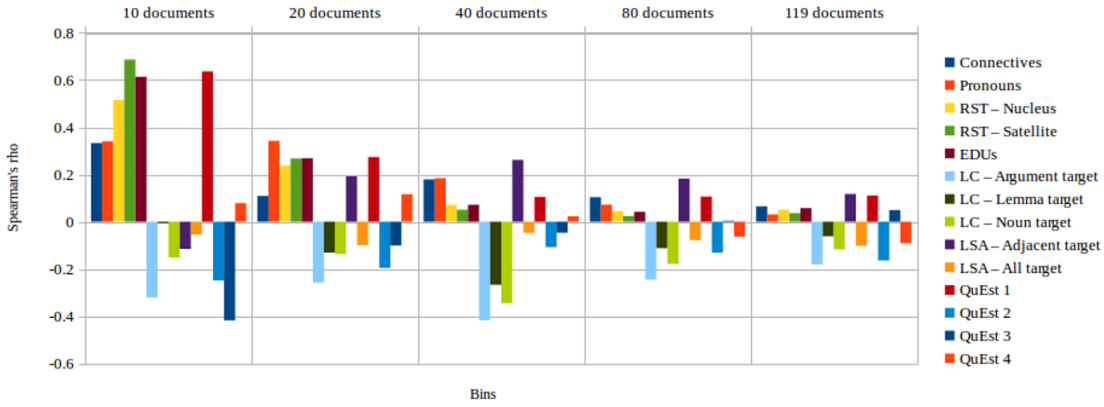

Fig. 4.2 Spearman's ρ correlation between target features and HTER values on the LIG corpus.

case of baseline features, only a feature that is known to perform very well for QE at sentence level (*QuEst* 1) achieves a high enough correlation score (above 0.6 for Pearson's *r* and Spearman's ρ), comparable to *RST- Satellite* and *EDUs*. All the other features achieved correlations of 0.4 or below. This provides evidence of how discourse-aware features differ from the baseline features that were inspired by sentence-level features.

It is worth emphasising that the experiments with different bins provide an evaluation of what would happen if the quality label used by evaluating the documents was very distinctive. Whilst when we use all the documents the majority of the features do not show high correlation, the situation for bins with higher variation is different. Moreover, the bins

also show the tails of the HTER distribution. Perhaps a system that aims to improve the prediction for outliers could benefit from such features.

Results for EN-FR documents from the Trace corpus (entire corpus, no bins) are shown in Figure 4.3 (Pearson's r correlation) and Figure 4.4 (Spearman's ρ correlation). Again, the results for both metrics are considerably similar (with some rare exceptions). In this case, the analysis was done in the source side only, and the QUEST++ features used were *QuEst* 5 to *QuEst* 17. For the analysis of English as source, the best feature is *QuEst* 5 with correlation scores below -0.4 for Pearson's r and below -0.5 for Spearman's ρ, followed by *LC - Argument source* (with almost 0.4 points for both correlation metrics). However, all discourse features showed correlations of above 0.2 or below -0.2 (with both Pearson's r and Spearman's ρ), higher than most QUEST++ features.

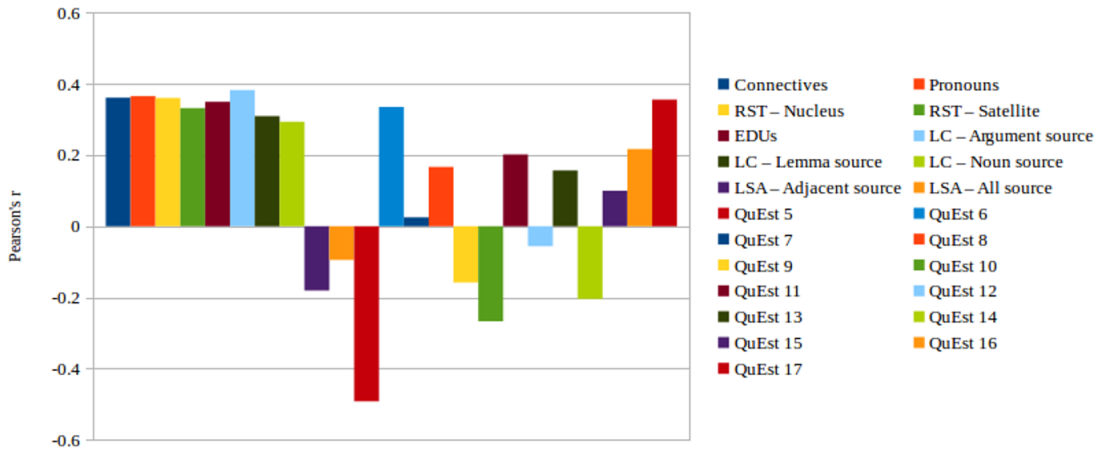

Fig. 4.3 Pearson's r correlation between target features and HTER values on the Trace corpus.

In order to better understand some of the discourse phenomena and the reasons behind their correlation with HTER, we conducted an analysis with the following features: number of pronouns, number of connectives and number of EDU breaks for the 10-document bin of the LIG corpus. Although these features do not correspond to all the best features identified in the previous section, they are the ones that are feasible to analyse manually or semi-automatically. The pronoun count achieved 0.34 points of Spearman's ρ correlation and 0.5 of Pearson's r correlation, but the p-values were higher than 0.05. This means that the correlation could be by chance. Pronouns were therefore analysed manually. Example (xvi) shows a case of problems with pronouns found in the LIG corpus, where MT is the machine translation and Post-Editing (PE) its post-edited version. In this example, there is a change in the pronoun "it" in the MT output, corrected to "he" in the post-edition.

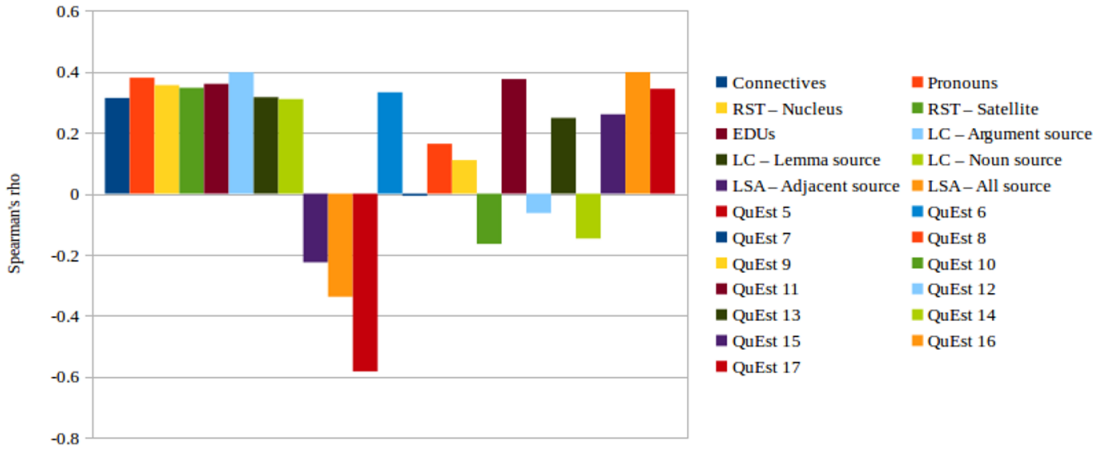

Fig. 4.4 Spearman's ρ correlation between target features and HTER values on the Trace corpus.

(xvi) **MT:** "Obviously, Gordon Brown wants to succeed Tony Blair as British prime minister. [...] Indeed, **it** has to renege on Blair's legacy, which, at least means promise to leave Britain for the Iraq war."
PE: "Obviously, Gordon Brown wants to succeed Tony Blair as British Prime Minister. [...] Indeed, **he** absolutely has to disavow Blair's legacy, which at least means promising to get Great Britain out of the Iraq war."

Example (xvii) presents a case where the pronoun "it" is removed in the post-edited version.

(xvii) **MT:** "It is the problem that **it** is the most urgent need to address: but for now, none of the main political parties has dared to touch it."
PE: "This is a problem that must be addressed immediately, but for now, none of the major political parties has dared to touch it."

Since the correlation between the number of pronouns against HTER was positive, the five documents with the highest HTER were manually evaluated looking for pronouns that were corrected from the MT version to the PE version. Figure 4.5 shows the total number of pronouns against the number of incorrect pronouns for the five documents. The number of incorrect pronouns is quite small compared to the total number of pronouns (proportionally, 23%, 10%, 16%, 33% and 34%, respectively in the five documents). This indicates that the high correlation showed between the number of pronouns and HTER was by chance in this corpus. However, it could also be an indication that the presence of pronouns led to sentences

that were more complicated and therefore more difficult to translate correctly (even if the pronouns themselves were correctly translated).

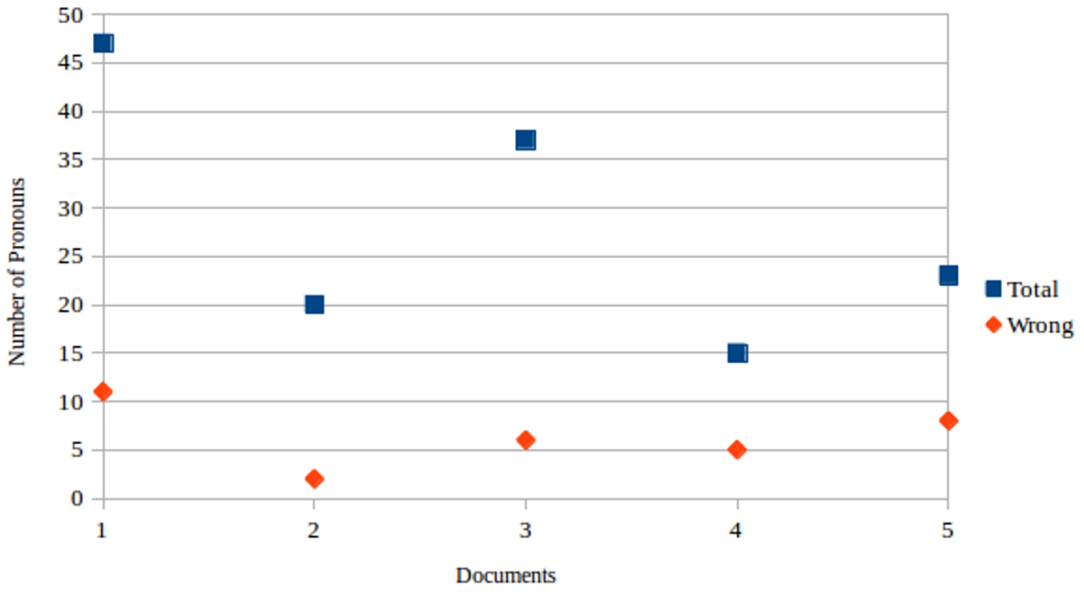

Fig. 4.5 Total number of pronouns and number of incorrectly translated pronouns for the top five documents in the LIG corpus.

Connectives were analysed in terms of numbers of connectives in the MT and PE versions and also the number per class, considering the classification in Pitler and Nenkova (2009): *Expansion*, *Contingency*, *Comparison*, *Temporal* and *non-discourse*. As in the case of pronouns, connectives showed a positive correlation with HTER (0.4 Pearson's r and 0.33 Spearman's ρ) and the p-values were also higher than 0.05. Figure 4.6 shows the number of connectives in the top five documents. As we can see, there is a change in the distribution of classes of connectives from the MT version to the PE version, i.e. the number of connectives in a given class changes from MT to PE. However, only document 4 showed significant changes. Therefore, it appears that the correlation between the number of connectives and HTER is by chance in this corpus.

In the case of EDUs, the p-values for the Pearson's r and Spearman's ρ correlation scores for the 10-document bin were below 0.05, meaning that the correlation is not by chance. Moreover, there we observed a change from the number of EDUs in the MT to the number of EDUs in the PE version. Therefore, we can infer that EDU breaks had an impact on the changes made to correct the documents, and thus on the MT quality of such documents.

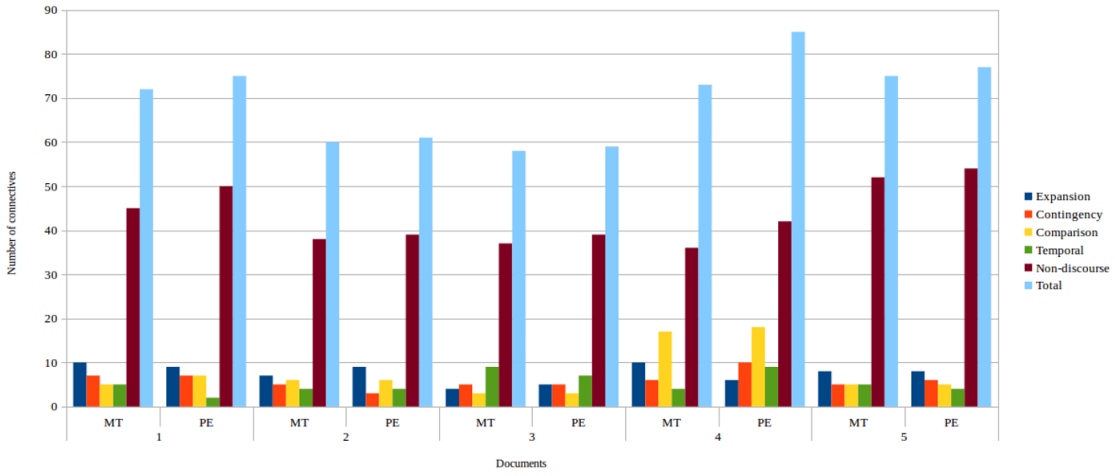

Fig. 4.6 Number of connectives in the MT and PE versions of the top five documents in the LIG corpus.

To avoid the bias of the top five documents, an additional analysis was done with 30 documents randomly selected from the 119 documents in the LIG corpus. We were interested in evaluating the impact of the same phenomena (number of pronouns, number of connectives and number of EDU breaks), but in a more general scenario. Figure 4.7 shows the percentage of incorrectly translated pronouns versus HTER figures. Although the distribution of percentages of incorrectly translated pronouns is different from the HTER distribution, the correlation between number of pronouns and HTER was quite high: 0.45 for Pearson's r (p-value $= 0.01$) and 0.31 for Spearman's ρ (p-value $= 0.1$). Therefore, we can conclude that there is a positive correlation between HTER scores and number of pronouns in this sample, and that it is not by chance.

For number of connectives, the correlation found was also high and significant: Pearson's r value of 0.52 (p-value $= 0.0$) and Spearman's ρ value of 0.48 (p-value = 0.0). The same was found for EDU breaks: the correlation found was 0.38 of Pearson's r (p-value = 0.04) and 0.44 of Spearman's ρ (p-value = 0.01). This means that the correlation between HTER values and number of EDU breaks is also not by chance.

4.6 Discussion

In this chapter we presented the different types of features used in our experiments with document-level QE. The document-aware features were an important extension for the QuEst++ tool (Section 4.1), although our main contribution was the proposal and develop-

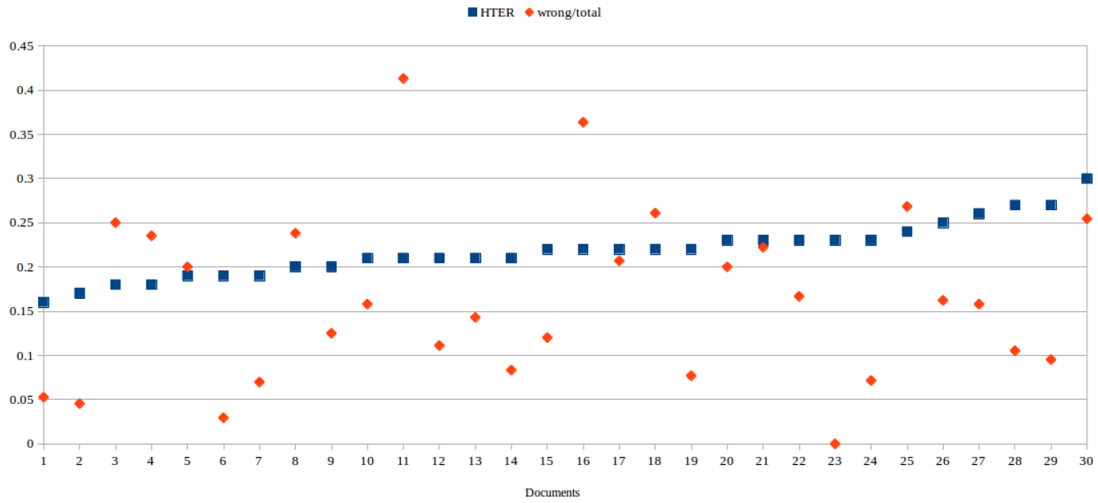

Fig. 4.7 HTER values versus percentage of incorrectly translated pronouns in a random sample of 30 documents from the LIG corpus.

ment of discourse-aware features, some of which were also included in QUEST++ (Section 4.2). Even though pseudo-reference features were also implemented, in the next chapters we focus on using consensus features (Section 4.4). Despite the fact that pseudo-reference-based features are considered to be a state-of-the-art resource for document-level QE, they are not completely reliable. Firstly, they work under the assumption that the pseudo-references have high quality. Secondly, they are not applicable for real-world scenarios, where MT outputs other than the output being evaluated are scarce. Consensus features have the same drawbacks and, therefore, they should be avoided in real-world scenarios. All features presented in this chapter are used to built the QE models presented in Chapters 5 and 6.

Discourse features appear to correlate with HTER scores and this correlation is often higher than the correlation presented by the basic, baseline QE features at document level (Section 4.5). Although HTER may not be the most reliable quality label for the task of QE (as we discuss in the next chapter), the correlation of discourse-aware features with task-based scores is promising. Therefore, we can conclude that discourse-aware features can be used as a source of information for QE at document level.

In the next chapter, we present our experiments with the features proposed in this chapter and using BLEU-style metrics as quality labels. We show results for three different datasets that include EN-PT, EN-DE, EN-ES, EN-FR, DE-EN, ES-EN and FR-EN language pairs.

Chapter 5

Document-level QE: Prediction

In this chapter we present and discuss document-level prediction using traditional MT automatic evaluation metrics (e.g. BLEU) as quality labels.[1] MT automatic evaluation metrics (as discussed in Chapter 2) have been widely used given that they require much less effort than manual evaluation and provide robust evaluation when the task is to compare different MT systems. For this reason, early work on document-level QE used such metrics as quality labels (Soricut and Echihabi, 2010; Scarton and Specia, 2014a). We extend such previous work and explore discourse information as features for document-level QE. Since BLEU-style metrics vary between 0 and 1 in a continuous scale, QE is addressed as a regression problem.

Section 5.1 describes the experimental settings common to all experiments performed in this chapter. Such settings include the feature sets used, the ML techniques used and a discussion about the evaluation metrics employed. Section 5.2 presents the document-level QE experiments with the FAPESP corpus, exploring single and multiple MT systems. BLEU, TER and METEOR are used as quality labels for the task. We consider two type of experiments: (i) with single systems, where all documents were translated by the same MT system and; (ii) with multiple MT systems explores the idea that documents are not translated by the same MT system. In Section 5.3, experiments document-level QE with different languages and mixed systems are presented. The data used is a selection of documents from WMT data for EN-DE, EN-ES, EN-FR, DE-EN, ES-EN and FR-EN. BLEU, TER and METEOR are also used as quality labels for the task. Section 5.4 presents our document-level QE experiments with the LIG corpus (FR-EN language pair) that also contains HTER

[1] Parts of this chapter have been previously published in peer-reviewed conferences: Scarton and Specia (2014a), Scarton et al. (2015), Scarton, Tan, and Specia (2015), Scarton and Specia (2016) and Scarton et al. (2016).

labels. Finally, Section 5.5 presents a discussion about the data distribution of the automatic evaluation metrics in all datasets.

5.1 Experimental Settings

Features

We conducted experiments using all the features described in Chapter 4 in all scenarios where they were available. As presented in Chapter 4, Most of the discourse-aware features are only available for the English language and, given the limitation of resources and tools, were only applied for English as the source language. We avoid applying NLP tools to machine translated documents, given that such tools assume well-formed documents as inputs.

Features are divided in different groups, following the classification presented on Chapter 4:

- QUEST++ baseline features (called hereafter QUEST-17): 17 QUEST++ baseline features (the same baseline features used for the WMT document-level shared task (Bojar et al., 2015, 2016b));
- Document-aware features (called hereafter QUEST-ALL): all QUEST++ features, apart from LC and LDA features;
- Language independent discourse-aware features (called hereafter SHALLOW): language independent discourse features include LSA, LC and LDA features;
- Word embeddings (called hereafter WE): word embeddings features for source and target documents. Although we classify such features as discourse-aware, they were not combined with SHALLOW features given their high dimensionality: they would probably obfuscate the use of the other SHALLOW features;
- Language dependent discourse-aware features (called hereafter DEEP): language dependent discourse features are pronouns, connectives and EDUs counts and RST tree information;
- Consensus features (called hereafter CONSENSUS): BLEU, TER and METEOR are used as consensus features. For each dataset a different number of pseudo-reference systems were available;
- All features (called hereafter ALL): a combination of all features.

Features are usually combined with the baseline features and cases where they are used alone are specified defined and justified.

ML models

We use two ML methods to build the regression models for document-level QE. Both methods are at the state-of-the-art of sentence-level QE and our work on document-level QE focuses on these methods.

SVM is a non-probabilistic model for supervised ML classification and regression. SVM is widely used in supervised ML problems, achieving good performance, particularly for applications where the number of features is considerably higher than the number of instances (Rogers and Girolami, 2012). Non-linear models can be built by using robust kernels. Kernels transform the data in order to make it classifiable by the linear model and, consequently, more easily modelled. We use the SVR algorithm available in the `scikit-learn` toolkit with RBF kernel and hyper-parameters (C, γ and ε) optimised via grid search.

GP is a probabilistic non-parametric model for supervised ML. GP generalises the Gaussian probability distribution, by dealing with the properties of a function instead of taking into account all the values the function outputs. Therefore, GP is a robust method and is computationally traceable (Rasmussen and Williams, 2006). This model has become more popular along the years among the NLP community. With GP models we also explore kernel combinations of features of different types. Our hypothesis is that features extracted from very distinct resources or using different techniques may benefit from the use of specific kernels for them. Therefore, we split features in three handcrafted (document and discourse-aware features), word embeddings and pseudo-references. Each of these sets, when combined in a single model, are addressed in different kernels by the GP models. We use the GPy toolkit[2] with RatQuad kernels and the optimisation of hyperparameters is done by maximising the model likelihood on the full training data.[3]

Evaluation

In the majority of previous work considering QE as a regression task, Mean Absolute Error (MAE) has been used as the evaluation metric (Callison-Burch et al., 2012; Bojar

[2]https://sheffieldml.github.io/GPy/

[3]We also experimented with Matern32, RBF and Exponential for GP. In our experiments, RatQuad showed consistently the best results.

et al., 2013, 2014, 2015). However, as Graham (2015) points out, MAE is not reliable to evaluate QE tasks, because it is highly sensitive to variance. This means that, if the predictions of a given QE model show high variance, it will lead to a high MAE, even though the distribution of the predictions follows the true labels distribution. Graham (2015) shows that such problem is very salient in datasets for QE at sentence level, which could mean that they are also extended to document-level QE, given that the two tasks are addressed similarly. In order to solve the issue, Graham (2015) suggests the use of the Pearson's r correlation coefficient as a metric of QE system evaluation.

Therefore, in the experiments with QE as a regression problem, we use Pearson's r as the main evaluation metric. MAE is still used as secondary metric since information about the variance of the data is also important. We believe that both metrics should be used in conjunction in order to reliably evaluate QE tasks.

Pearson's r correlation coefficients varies between -1 and 1, where -1 is the maximum value of negative correlation, whilst 1 is the maximum value of positive correlation. There is no consensus on a threshold on the r coefficient that defines when a correlation should be considered weak, moderate or strong. For the purposes of our experiments we considered the thresholds in Table 5.1.[4]

r	Correlation type
-1.0 to -0.5 and 0.5 to 1.0	strong correlation
-0.5 to -0.3 and 0.3 to 0.5	moderate correlation
-0.3 to -0.1 and 0.1 to 0.3	weak correlation
-0.1 to 0.1	no correlation

Table 5.1 Thresholds on Pearson's r correlation coefficients used in our experiments.

Pearson's r calculation also provide a p-value, indicating whether or not the correlation is statistically significant. We assume p-value < 0.05 indicates statistical significance (95% of confidence).

MAE is calculated using Equation 5.1 where $H(s_i)$ is the predicted score, $V(s_i)$ is the true score and N is the number of data points in the test set.

$$MAE = \frac{\sum_{i=1}^{n} |H(s_i) - V(s_i)|}{N} \tag{5.1}$$

Finally, when comparing two different systems, we apply Williams's significance test (as discussed by Graham (2015)) with significant values having p-value < 0.05. Therefore

[4]Explorable.com (May 2, 2009). Statistical Correlation. Retrieved Jul 05, 2016 from Explorable.com: https://explorable.com/statistical-correlation.

everytime that we compare two different QE systems and state that one system showed significant higher correlation than the other system, we will be referring to the use of this test.

Baselines

Two baselines are used, depending on the evaluation metric set for a given experiment. The first baseline (MEAN) uses the mean value of all labels in the training set as the predicted value of all test set instances. This is a strong baseline because of the tendency showed by the document-level data where all values are close to the mean value (see Section 5.5). However, MEAN cannot be used to evaluate Pearson's r results, since its variance will always be zero. MEAN is then used to evaluate the documents in terms of **performance gain** or **error reduction**. Equation 5.2 shows how we calculated this performance gain, where MAE_{mean} is the value of the MAE for the mean baseline and $MAE_{prediction}$ is the MAE of the QE model under investigation.

$$gain = \frac{MAE_{mean} - MAE_{prediction}}{MAE_{mean}} * 100 \tag{5.2}$$

Therefore, we use a random baseline (RANDOM) to compare the results of Pearson's r correlation coefficients. In this baseline, for each test set instance, a label from the training set is randomly selected to be the label of the test instance.

An alternative baseline could be the mean of sentence-level predictions. However, we could not train sentence-level QE systems for all language pairs using a reliable quality label (such as HTER). Therefore, we decided not to include such a baseline.

5.2 Experiments with a Large Corpus: FAPESP Data

The first experiments aim at document-level QE focusing on building QE models for English (EN) into Brazilian Portuguese (BP) MT.[5] Following (Soricut and Echihabi, 2010), we considered BLEU-style metrics as quality labels. Document-aware, discourse-aware and pseudo-references and are applied.

Corpus **FAPESP** contains 2,823 English-Brazilian Portuguese (EN-BP) documents extracted from a scientific Brazilian news journal (FAPESP)[6] (Aziz and Specia, 2011). Each

[5]This is an extension of our work in Scarton and Specia (2014a)

[6]http://revistapesquisa.fapesp.br

article covers one particular scientific news topic. The corpus was randomly divided into 60% (1,694 documents) for training a baseline MOSES[7] statistical phrase-based MT system (Koehn et al., 2007) (with 20 documents as development set); and 40% (1,128 documents) for testing the SMT system, which generated translations for QE training (60%: 677 documents) and test (40%: 451 documents). In addition, two external MT systems were used to translate the test set: SYSTRAN[8] – a rule-based system – and Google Translate (GOOGLE)[9] (the latter is used as pseudo-reference for the other MT systems given that its overall BLEU score was better than that of the others).

Features In the experiments reported in this section we used the following feature sets: QUEST-17, QUEST-ALL, QUEST-17+SHALLOW, QUEST-ALL+SHALLOW, QUEST-17+WE, QUEST-17+PSEUDO and ALL. Although English is the source language, the documents of this corpus had several special tags and were considerably large, which made the use of the discourse parser infeasible.

Quality labels The automatic metrics selected for quality labelling and prediction are BLEU, TER and METEOR. The Asiya toolkit was used to calculate all metrics.

Method Two sets of experiments were conducted. First (Section 5.2.1), we consider the outputs of the FAPESP corpus of MOSES and SYSTRAN separately, using as training and test sets the output of each system individually, with GOOGLE translations as pseudo-reference for the other two systems. In this case, the quality of GOOGLE is known to be better than that of MOSES and SYSTRAN (for the purposes of this experiment, the GOOGLE pseudo-reference features - BLEU, TER and METEOR - are called PSEUDO). The second set of experiments (Section 5.2.2) considers, for the FAPESP corpus, the combination of the output of all systems (MIXED), and we use the concept of **consensus**: for each system the other two are used as pseudo-references. SVR and GP are used to generate the QE models.

5.2.1 MT System-specific Models

The results for the prediction of BLEU, TER and METEOR for MOSES using SVR and GP models (in terms of Pearson's *r*) are shown in Table 5.2, whilst results for SYSTRAN are shown in Table 5.3.

[7] http://www.statmt.org/moses/?n=moses.baseline
[8] http://www.systransoft.com/
[9] https://translate.google.co.uk

The best results in terms of Pearson's *r* correlation for MOSES and SYSTRAN were obtained with pseudo-references and SVR models for all labels (BLEU, TER and METEOR). For MOSES, the use of QUEST-ALL features showed significant improvement over QUEST-17 in all cases, whilst for SYSTRAN only three cases showed improvements: SVR predicting TER and GP predicting TER and METEOR. For SYSTRAN, ALL feature set showed significant improvements over QUEST-17 when predicting TER and METEOR using GP, while for MOSES GP models with the ALL feature set showed significant improvements over QUEST-17. Although GP models used different kernels for the different types of features, the best result with pseudo-references were achieved with SVR. On the other hand, kernel combination seems to help when using ALL features, since GP models built with one kernel for representing QUEST-ALL+SHALLOW, one for WE and one for PSEUDO performs better than the SVR model with only one kernel. The use of SHALLOW and WE features did not improve over QUEST-17 in both cases. The baseline results are considerably high in both MOSES and SYSTRAN experiments and this may explain why the addition of more features does not necessarily improve the results. Finally, all predicted values performed better than the RANDOM baseline.

	BLEU		TER		METEOR	
	SVR	GP	SVR	GP	SVR	GP
QUEST-17	0.690	0.688	0.625	0.607	0.667	0.666
QUEST-ALL	0.725	0.717	0.687	0.657	0.714	0.699
QUEST-17+SHALLOW	0.607	0.564	0.564	0.496	0.609	0.547
QUEST-ALL+SHALLOW	0.684	0.662	0.640	0.602	0.677	0.649
QUEST-17+WE	0.517	0.664	0.463	0.599	0.516	0.641
QUEST-17+PSEUDO	**0.857**	0.848	**0.806**	0.786	**0.846**	0.832
ALL	0.686	0.842	0.623	0.780	0.687	0.820
RANDOM	0.004*		0.025*		0.013*	

Table 5.2 Results for MOSES system in terms of Pearson's *r* correlation. * indicates results that did not show significant Pearson's *r* correlation with p-value < 0.05. The best systems are highlighted in bold (William's significance test with p-value < 0.05).

In terms of performance gains (Equation 5.2), Figures 5.1 and 5.2 show the results for MOSES and SYSTRAN considering SVR and GP models and all feature sets combination. QE systems built with QUEST-17+PSEUDO showed the best performance gain varying between 35% and 50%. The performance of PSEUDO features can be explained by the quality of the pseudo-reference employed. GOOGLE showed +0.09 BLEU points when compared to MOSES and +0.18 BLEU points when compared to SYSTRAN. Therefore, GOOGLE is a more reliable system and the pseudo-reference-based features are competitive.

	BLEU		TER		METEOR	
	SVR	GP	SVR	GP	SVR	GP
QUEST-17	0.598	0.567	0.489	0.488	0.495	0.492
QUEST-ALL	0.541	0.490	0.508	0.491	0.488	0.494
QUEST-17+SHALLOW	0.124	0.123	0.154	0.274	0.139	0.152
QUEST-ALL+SHALLOW	0.439	0.406	0.427	0.438	0.449	0.426
QUEST-17+WE	0.296	0.237	0.323	0.369	0.317	0.289
QUEST-17+PSEUDO	**0.627**	0.613	**0.703**	0.625	**0.653**	0.619
ALL	0.445	0.566	0.414	0.602	0.435	0.571
RANDOM	0.003*		0.000*		0.032*	

Table 5.3 Results for SYSTRAN system in terms of Pearson's r correlation. * indicates results that did not show significant Pearson's r correlation with p-value < 0.05. The best systems are highlighted in bold (William's significance test with p-value < 0.05).

On the other hand, ALL features did not show competitive results and only showed significant gains when the models were built with GP. In this case, the use of different kernels for different feature sets seems to be the best choice (the same happened for QUEST-17+WE). In general, all systems showed performance gain over the MAE_{mean}. However, it is worth noticing that the MAE_{mean} was 0.058 for BLEU, 0.074 for TER and 0.057 for METEOR, which are small error values for labels that vary between 0 and 1. An extended discussion about this topic is provided in Section 5.5.

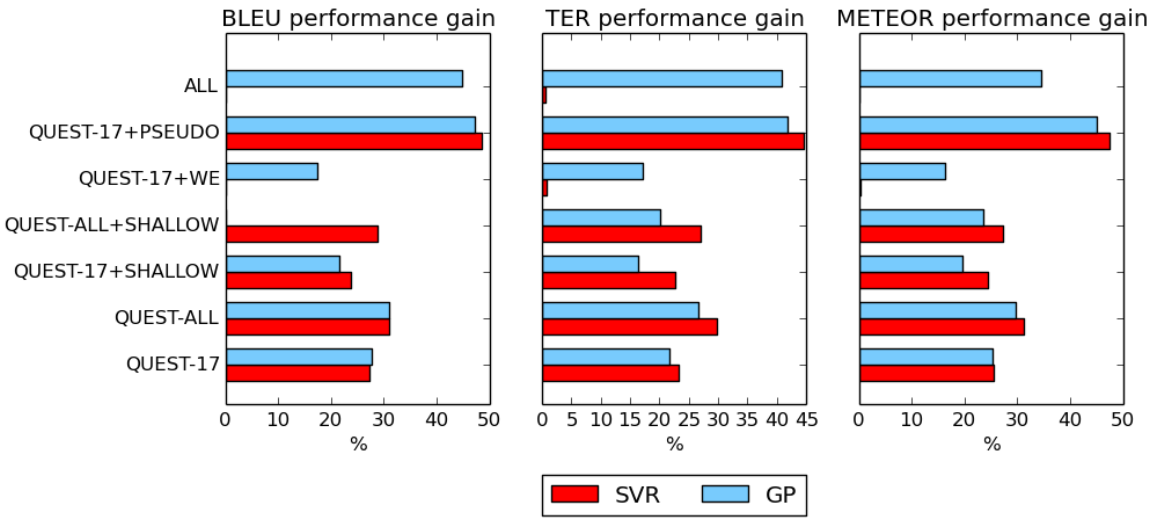

Fig. 5.1 Performance gains in terms of MAE of the QE models for MOSES documents.

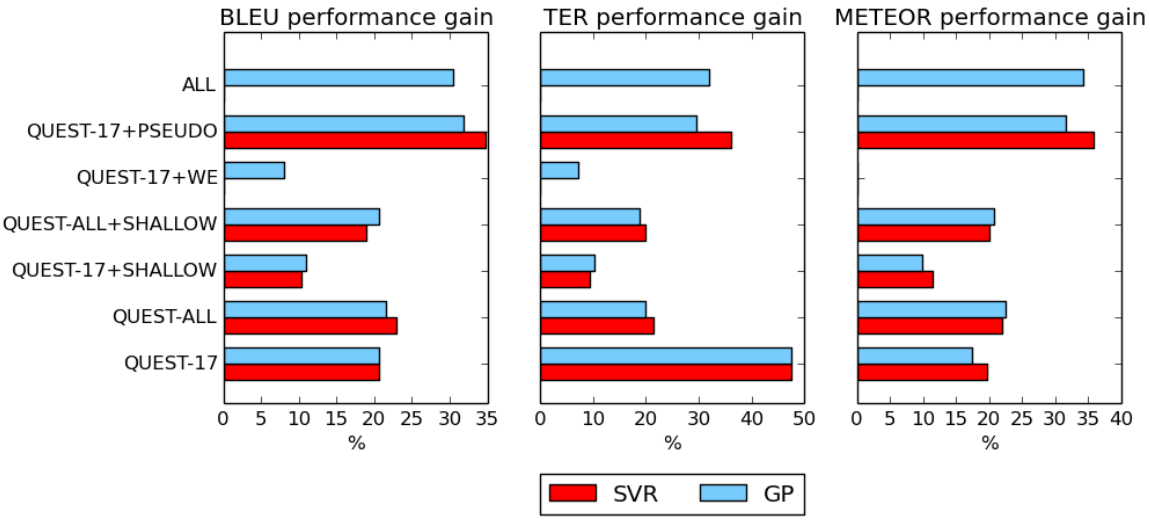

Fig. 5.2 Performance gains in terms of MAE of the QE models for SYSTRAN documents.

5.2.2 MT Multiple-systems Models

In this experiment, the output of MOSES, SYSTRAN and GOOGLE is mixed and, for each system being used, the CONSENSUS features are extracted by using the other two systems as pseudo-references. CONSENSUS features led to the best results in Table 5.4 (as the PSEUDO features in previous section), although the best models differ: for BLEU and TER the best system was built with SVR, whilst for METEOR the best system was built with GP. Similar to the system-specific experiments, the good performance of CONSENSUS features could be related to the fact that the MT systems are significat different in terms of quality (at least +0.86 BLEU points between adjacent systems). Therefore, when GOOGLE is a pseudo-reference, it works as an upper bound, whilst SYSTRAN would work as a lower bound (similar to the work presented by Louis and Nenkova (2013)). Again, the use of QUEST-ALL features outperform QUEST-17, whilst the use of the SHALLOW feature set does not outperform the baseline. WE features are better than QUEST-17 only when GP models (with two kernels) are used for BLEU and METEOR (for TER the model built with WE feature set is not significantly different from QUEST-17). For the ALL feature set, GP models show a significant improvement over SVR models, which is in accordance with our hypothesis that GP models with different kernel combinations are more robust because they treat different feature sets separately.

Figure 5.3 shows the performance gains for the MIXED experiment. QE systems built with QUEST-17+CONSENSUS showed the best performance gains varying between 50%

	BLEU		TER		METEOR	
	SVR	GP	SVR	GP	SVR	GP
QUEST-17	0.565	0.621	0.559	0.566	0.565	0.598
QUEST-ALL	0.620	0.660	0.598	0.594	0.623	0.661
QUEST-17+SHALLOW	0.486	0.539	0.437	0.440	0.460	0.516
QUEST-ALL+SHALLOW	0.572	0.599	0.546	0.484	0.577	0.538
QUEST-17+WE	0.443	0.635	0.423	0.572	0.414	0.622
QUEST-17+CONSENSUS	**0.912**	0.911	**0.855**	0.847	0.885	**0.890**
ALL	0.678	0.903	0.614	0.835	0.661	0.882
RANDOM	0.009*		0.100*		0.011*	

Table 5.4 Results for MIXED in terms of Pearson's r correlation. * indicates results that did not show significant Pearson's r correlation with p-value < 0.05. The best systems are highlighted in bold (William's significance test with p-value < 0.05).

and 70% over the MAE_{mean} baseline, agreeing with the results showed for Pearson's r correlation. ALL features also showed good results in terms of performance gains when GP models were considered. Again, the use of different kernels for different feature sets appears to be important when dealing with different types of features (mainly WE, where QUEST-17+WE only showed performance gains for models built with GP). In general, all systems showed performance gains with respect to MAE_{mean}.

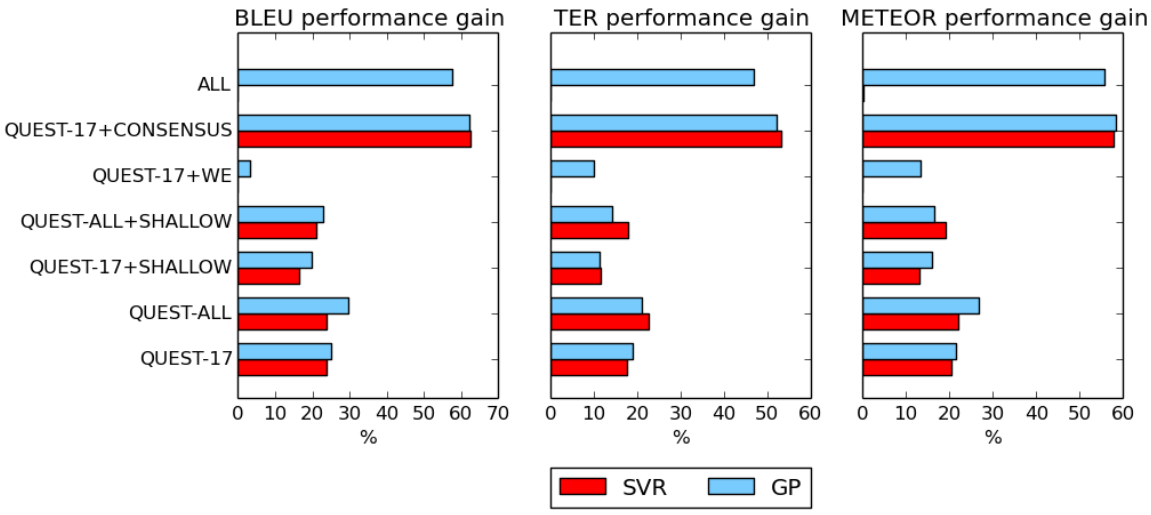

Fig. 5.3 Performance gains in terms of MAE of the QE models for MIXED documents.

5.3 Experiments with Multiple Language Pairs: WMT Data

In this section we extend the work done for the FAPESP corpus to several languages. The translations come from different MT systems and, therefore, the settings are comparable to those in the MIXED scenario.

Corpus Our **WMT** corpus contains 474 news documents in six language pairs (EN-DE, EN-ES, EN-FR, DE-EN, ES-EN and FR-EN) from the WMT translation shared task corpora from editions 2008 to 2013 (Callison-Burch et al., 2008, 2009, 2010, 2011, 2012; Bojar et al., 2013). The corpus was randomly divided into 70% (332 documents) and 30% (142 documents) for training and testing of the QE models. For each source document in the collections, a translation from the set of participating MT systems was randomly selected. Therefore, documents have translations from different MT systems. The participating MT systems for the translation shared task include SMT, Rule-based MT (RBMT) and hybrid systems (Table 5.5 shows the overall performance of the participating systems per year - in terms of BLEU).

	MIN	MAX	MEAN
EN-DE	0.10	0.21	0.15
EN-ES	0.15	0.34	0.26
EN-FR	0.13	0.32	0.24
DE-EN	0.07	0.30	0.20
ES-EN	0.16	0.35	0.27
FR-EN	0.14	0.33	0.26

Table 5.5 Overall performance, in terms of BLEU, of MT systems submitted to WMT shared tasks from EN into DE, ES and FR and from these languages into EN (values are calculated over all systems submitted for the WMT editions between 2008 and 2013).

Features In the experiments reported in this section we use the following feature sets: QUEST-17, QUEST-ALL, QUEST-17+SHALLOW, QUEST-ALL+SHALLOW, QUEST-17+WE, QUEST-17+CONSENSUS and ALL for all language pairs. CONSENSUS features are extracted using, for each translation from a given MT system, the other systems as pseudo-references. For translations from English, QUEST-17+DEEP is also explored.

Quality labels The automatic metrics selected for quality labelling and prediction are BLEU, TER and METEOR. The Asiya toolkit was used to calculate these metrics.

Method SVR and GP are used to generate the QE models.

EN-DE Differently from the FAPESP corpus, the majority of the QE models built did not show significant correlation against the true labels. In fact, only models with QUEST-17+WE, QUEST-17+CONSENSUS and ALL show significant correlation scores. In the WE case, all results show strong correlation (Pearson's $r > 0.5$), while in the CONSENSUS case, two results show strong correlation, three show moderate correlation (Pearson's $0.3 < r < 0.5$) and one shows weak correlation (Pearson's $0.1 < r < 0.3$). ALL feature set shows strong correlation scores for all models and, when using GP with different kernels for different types of features, they achieve the best correlation scores. In all cases using more than one kernel for GP (QUEST-17+WE, QUEST-17+CONSENSUS, ALL), GP models outperforms their SVR counterparts (the only exception are the models predicting METEOR with QUEST-17+WE feature set).

	BLEU		TER		METEOR	
	SVR	GP	SVR	GP	SVR	GP
QUEST-17	0.191*	0.162*	0.047*	0.105*	0.024*	0.076*
QUEST-ALL	0.074*	0.010*	0.115*	0.006*	0.006*	0.048*
QUEST-17+SHALLOW	0.052*	0.005*	0.069*	0.061*	0.000*	0.064*
QUEST-ALL+SHALLOW	0.072*	0.014*	0.021*	0.018*	0.046*	0.027*
QUEST-17+DEEP	0.015*	0.024*	0.018*	0.034*	0.005*	0.020*
QUEST-ALL+DEEP	0.030*	0.021*	0.010*	0.042*	0.025*	0.060*
QUEST-ALL+SHALLOW+DEEP	0.051*	0.010*	0.003*	0.028*	0.016*	0.027*
QUEST-17+WE	0.512	0.526	0.519	0.522	0.725	0.622
QUEST-17+CONSENSUS	0.389	0.531	0.324	0.446	0.256	0.728
ALL	0.534	**0.666**	0.531	**0.614**	0.730	**0.795**
RANDOM	0.263		0.155*		0.034*	

Table 5.6 Results for WMT EN-DE in terms of Pearson's r correlation. * indicates results that did not show significant Pearson's r correlation with p-value < 0.05. The best systems are highlighted in bold (William's significance test with p-value < 0.05).

Regarding performance gains, Figure 5.4 shows that the best models showed improvements between 25% and 30% (being the best results achieved with ALL features and GP models for all labels). Discourse features (both SHALLOW and DEEP) show no improvements or only marginal improvements. It is worth mentioning that the MT systems for this language pair showed very low performance (in average, 0.15 of BLEU - Table 5.5), which may impact the peformance of the QE systems (a more detailed analysis over the data is provided in Section 5.5).

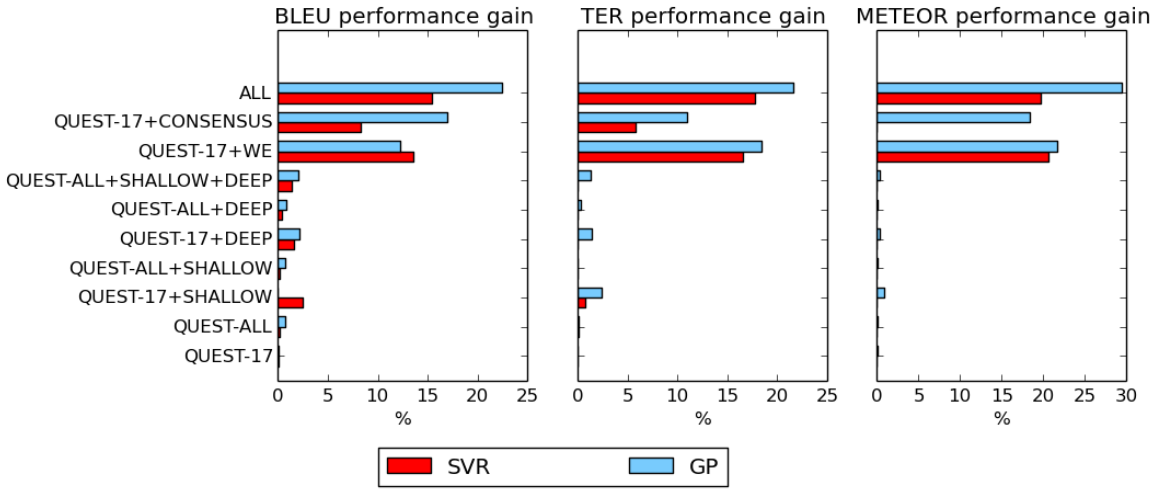

Fig. 5.4 Performance gains in terms of MAE of the QE models for WMT EN-DE documents.

EN-ES In contrast to EN-DE, Table 5.7 shows that all systems achieved significant Pearson's r correlations scores (apart from RANDOM). However, results for QUEST-17, QUEST-ALL, QUEST-17+SHALLOW, QUEST-ALL+SHALLOW, and QUEST-17+DEEP show weak correlation scores (Pearson's $r < 0.3$). The only exception is QUEST-ALL with a GP model predicting METEOR that shows 0.337 of Pearson's r correlation (a moderate correlation). The highest Pearson's r scores are obtained by the models built with QUEST-17+WE (with GP) and ALL (with both GP and SVR) feature sets for all metrics (according to William's test these systems do not show significant difference). For the WE feature set, the models with GP and two kernels outperformed the versions with SVR and one kernel. The same did not happened for CONSENSUS features, where the GP models did not outperform the SVR models (in fact, there is no significant differences between GP and SVR models in this case).

For EN-ES (Figure 5.5), the best results for performance gains were below 25% (the best systems being the ones with WE features and the GP model). Systems built with SHALLOW and DEEP features showed marginal performance gains, below 10%. The MT systems submitted show one of the highest averaged performances (0.26 averaged BLEU - Table 5.5). However, in terms of performance gain, the values are not far from the EN-DE case where the MT systems showed much lower performance.

EN-FR Similarly to EN-DE, the QUEST-17, QUEST-ALL, QUEST-17+SHALLOW, QUEST-ALL+SHALLOW, and QUEST-17+DEEP results did not show significant Pearson's

	BLEU		TER		METEOR	
	SVR	GP	SVR	GP	SVR	GP
QUEST-17	0.204	0.181	0.258	0.226	0.240	0.206
QUEST-ALL	0.212	0.290	0.275	0.380	0.243	0.337
QUEST-17+SHALLOW	0.245	0.187	0.258	0.244	0.268	0.212
QUEST-ALL+SHALLOW	0.161	0.270	0.105	0.102	0.207	0.319
QUEST-17+DEEP	0.201	0.218	0.265	0.229	0.252	0.238
QUEST-ALL+DEEP	0.238	0.286	0.282	0.377	0.263	0.334
QUEST-ALL+SHALLOW+DEEP	0.202	0.274	0.239	0.362	0.236	0.326
QUEST-17+WE	0.495	**0.511**	0.528	**0.551**	0.549	**0.559**
QUEST-17+CONSENSUS	0.316	0.296	0.340	0.341	0.272	0.260
ALL	**0.508**	**0.509**	**0.532**	**0.555**	**0.552**	**0.558**
RANDOM	0.090*		0.019*		0.043*	

Table 5.7 Results for WMT EN-ES in terms of Pearson's r correlation. * indicates results that did not show significant Pearson's r correlation with p-value < 0.05. The best systems are highlighted in bold (William's significance test with p-value < 0.05).

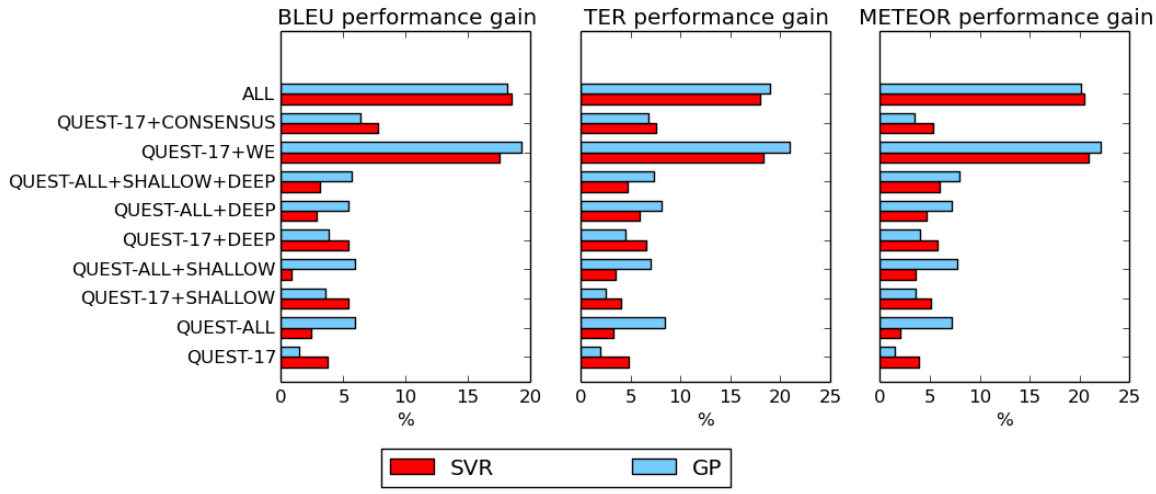

Fig. 5.5 Performance gains in terms of MAE of the QE models for WMT EN-ES documents.

r correlation against true labels (Table 5.8). The highest correlation scores for this language pair were found by using ALL features, with strong significant correlation for SVR models. WE and CONSENSUS combined with QUEST-17 also achieved strong correlation scores for some cases. Differently from EN-DE, the use of multiple kernels in the GP models did not outperform the SVR versions with a single kernel.

Figure 5.6 shows the performance gains for the models predicting BLEU, TER and METEOR for EN-FR language pair. The best gains, achieved by ALL features models, are below 18%. Discourse-aware features did not achieve significant improvements (all below

	BLEU		TER		METEOR	
	SVR	GP	SVR	GP	SVR	GP
QUEST-17	0.084*	0.012*	0.094*	0.037*	0.135*	0.098*
QUEST-ALL	0.009*	0.040*	0.066*	0.024*	0.003*	0.023*
QUEST-17+SHALLOW	0.048*	0.037*	0.061*	0.075*	0.108*	0.095*
QUEST-ALL+SHALLOW	0.008*	0.034*	0.046*	0.000*	0.021*	0.015*
QUEST-17+DEEP	0.111*	0.028*	0.055*	0.091*	0.074*	0.079*
QUEST-ALL+DEEP	0.008*	0.034*	0.046*	0.000*	0.021*	0.015*
QUEST-ALL+SHALLOW+DEEP	0.039*	0.016*	0.075*	0.035*	0.006*	0.016*
QUEST-17+WE	0.507	0.451	0.495	0.480	0.627	0.594
QUEST-17+CONSENSUS	0.431	0.269	0.352	0.283	0.541	0.458
ALL	**0.531**	0.430	**0.526**	0.475	**0.651**	0.585
RANDOM	0.091*		0.031*		0.117*	

Table 5.8 Results for WMT EN-FR in terms of Pearson's r correlation. * indicates results that did not show significant Pearson's r correlation with p-value < 0.05. The best systems are highlighted in bold (William's significance test with p-value < 0.05).

2% of gains). Although the MT systems for this language pair showed high performance (0.24 of averaged BLEU - Table 5.5), results for both performance gain and Pearson's r correlation are not better than for EN-DE.

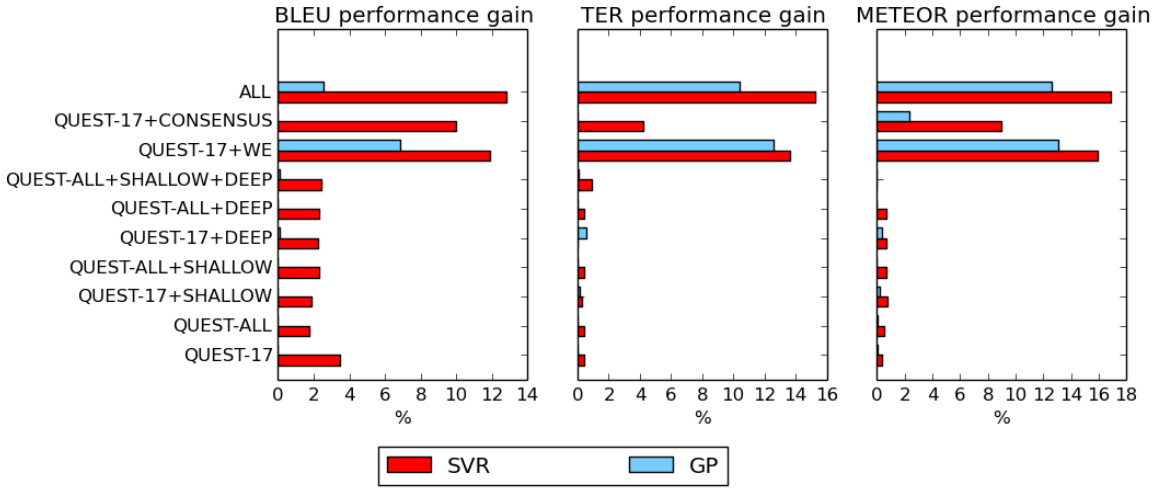

Fig. 5.6 Performance gains in terms of MAE of the QE models for WMT EN-FR documents.

DE-EN Table 5.9 shows that a significant Pearson's r correlation was found for all models in this language pair. The highest correlation scores were achieved by the model with the ALL feature set and SVR for BLEU, the ALL feature set and both ML models for TER and

the ALL and QUEST-17+WE feature sets with SVR for METEOR. For BLEU and TER the best results feature a strong correlation, whilst for METEOR the correlation is moderate. The use of different kernels in the GP models did not offer consistent improvements over the SVR models.

	BLEU		TER		METEOR	
	SVR	GP	SVR	GP	SVR	GP
QUEST-17	0.224	0.252	0.221	0.259	0.206	0.237
QUEST-ALL	0.250	0.285	0.354	0.397	0.307	0.313
QUEST-17+SHALLOW	0.328	0.372	0.224	0.293	0.289	0.367
QUEST-ALL+SHALLOW	0.280	0.331	0.376	0.392	0.352	0.323
QUEST-17+WE	0.483	0.473	0.563	0.553	**0.462**	0.483
QUEST-17+CONSENSUS	0.257	0.259	0.385	0.438	0.277	0.185
ALL	**0.508**	0.446	**0.624**	**0.630**	**0.491**	0.360
RANDOM	0.022		0.033*		0.142*	

Table 5.9 Results for WMT DE-EN in terms of Pearson's *r* correlation. * indicates results that did not show significant Pearson's *r* correlation with p-value < 0.05. The best systems are highlighted in bold (William's significance test with p-value < 0.05).

Figure 5.7 shows the results in terms of performance gains for DE-EN. The best gains were lower than 20% for BLEU and METEOR and lower 30% for TER and were achieved by models built with the ALL feature set and SVR. The overall quality of DE-EN MT systems was low (0.20 of averaged BLEU - Table 5.5) when compared to other language pairs. It is important noticing that, for this language pair, the best systems in terms of performance gain are built with different ML models for different labels (for BLEU and TER, the best model was built with GP, whilst for METEOR the best model was built with SVR).

ES-EN Apart from the two systems using the baseline features along with SVR models for predicting BLEU and TER, all other cases led to significant Pearson's *r* correlation scores (Table 5.10). The highest correlation scores were achieved with the use of ALL feature set, with no significant difference between GP and SVR models. For ALL feature set models, correlation scores for BLEU and METEOR are moderate, whilst correlation scores for TER are strong. The use of different kernels in the GP models for QUEST-17+CONSENSUS shows improvements over the SVR models only when predicting TER and METEOR. No significant improvements were shown by GP models over SVR models with the QUEST+WE feature set.

Regarding performance gains, Figure 5.8 shows that the best gains (for models built with ALL features) have performance gains of up to 16% for all automatic metrics. Discourse

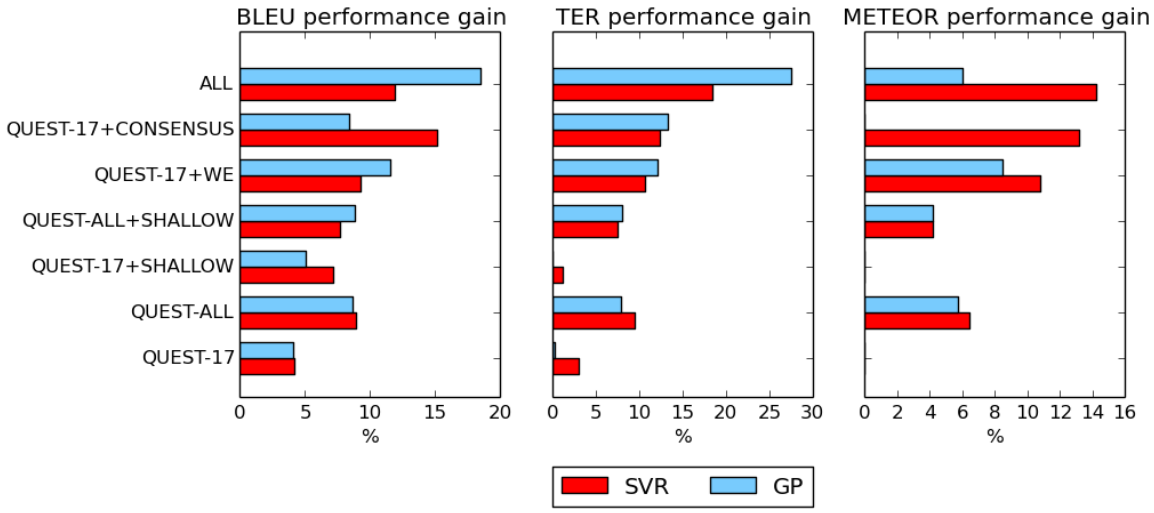

Fig. 5.7 Performance gains in terms of MAE of the QE models for WMT DE-EN documents.

	BLEU		TER		METEOR	
	SVR	GP	SVR	GP	SVR	GP
QUEST-17	0.160*	0.168	0.146*	0.200	0.195	0.168
QUEST-ALL	0.175	0.203	0.246	0.275	0.168	0.179
QUEST-17+SHALLOW	0.174	0.175	0.269	0.231	0.289	0.230
QUEST-ALL+SHALLOW	0.182	0.235	0.268	0.296	0.244	0.254
QUEST-17+WE	0.376	0.388	0.446	0.450	0.315	0.315
QUEST-17+CONSENSUS	0.323	0.348	0.304	0.331	0.245	0.364
ALL	**0.425**	**0.430**	**0.505**	**0.515**	**0.379**	**0.393**
RANDOM	0.103*		0.093*		0.054*	

Table 5.10 Results for WMT ES-EN in terms of Pearson's r correlation. * indicates results that did not show significant Pearson's r correlation with p-value < 0.05. The best systems are highlighted in bold (William's significance test with p-value < 0.05).

features did not yield considerable improvements. Similarly to the DE-EN case, for ES-EN, there was also a difference between the ML techniques applied in the best systems (in terms of performance gains). For BLEU and METEOR, the best systems were built by using SVR, whilst for TER the best system was built with GP. Although this language pair had the best MT systems in terms of averaged BLEU (0.27 - Table 5.5), the performance gains and Pearson's r scores are not higher than for other language pairs.

FR-EN Table 5.11 shows that, with the exceptions of GP models predicting BLEU and METEOR with QUEST-17+CONSENSUS features and predicting METEOR with ALL features, all cases showed significant Pearson's r correlation scores. The highest correlation

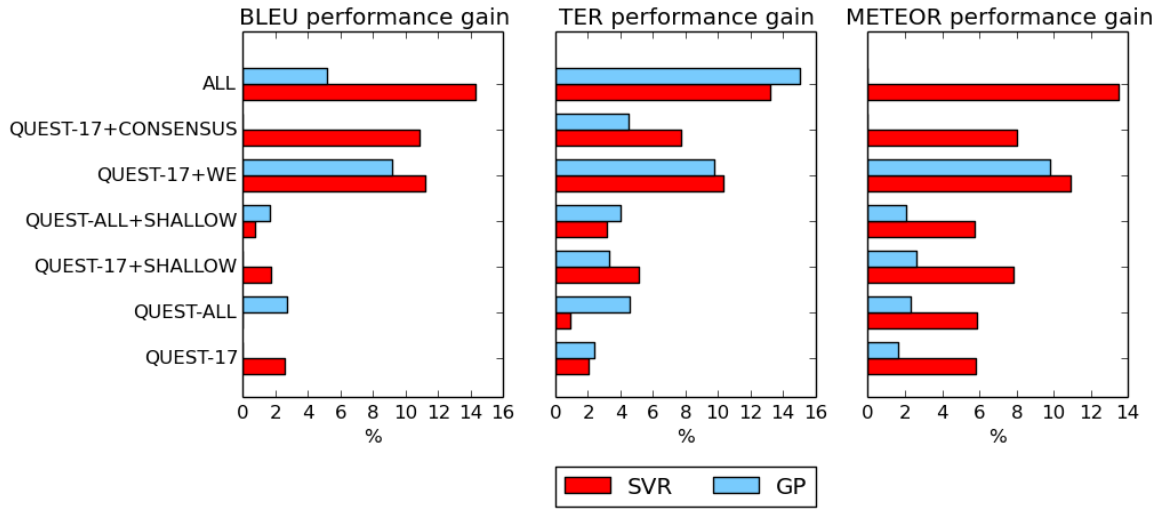

Fig. 5.8 Performance gain in terms of MAE of the QE models for WMT ES-EN documents.

scores for FR-EN were achieved by models built with the ALL feature set and SVR. However, when predicting METEOR, the model built with the ALL feature set and SVR was not significantly different from the model built with QUEST-17+CONSENSUS and SVR (according to William's test). For this language pair, the use of different kernels and GP models did not outperform the SVR models with one kernel only.

	BLEU		TER		METEOR	
	SVR	GP	SVR	GP	SVR	GP
QUEST-17	0.227	0.172	0.176	0.145	0.223	0.135
QUEST-ALL	0.288	0.250	0.305	0.247	0.320	0.270
QUEST-17+SHALLOW	0.266	0.215	0.285	0.229	0.242	0.143
QUEST-ALL+SHALLOW	0.317	0.308	0.327	0.321	0.315	0.323
QUEST-17+WE	0.405	0.432	0.497	0.485	0.410	0.410
QUEST-17+CONSENSUS	0.412	0.135*	0.358	0.308	**0.442**	0.081*
ALL	**0.457**	0.325	**0.544**	0.517	**0.463**	0.062*
RANDOM	0.132*		0.073*		0.144*	

Table 5.11 Results for WMT FR-EN in terms of Pearson's r correlation. * indicates results that did not show significant Pearson's r correlation with p-value < 0.05. The best systems are highlighted in bold (William's significance test with p-value < 0.05).

Figure 5.9 shows the performance gains for FR-EN. QE models with the ALL feature set achieved the best gains, with up to 16% of gains for all automatic metrics. SHALLOW features showed no considerable gains over the MAE_{mean} baseline. The MT systems for this

language pair also showed high performance (0.26 of averaged BLEU - Table 5.5), when compared to EN-DE and DE-EN for example.

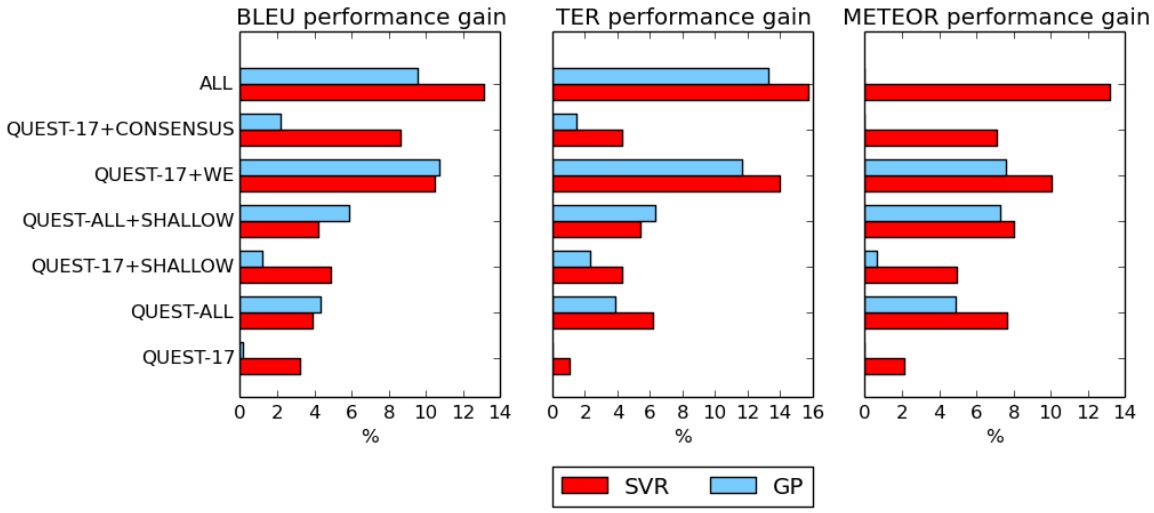

Fig. 5.9 Performance gains in terms of MAE of the QE models for WMT FR-EN documents.

Summary for all language pairs The results show a high variation among the different language pairs. For the majority of language pairs, the best results were achieved with the use of ALL features, although the best ML approach varied across different settings. Also, for the majority of the cases, the use of CONSENSUS or WE features improved over the QUEST-17 models. Discourse-aware features (SHALLOW and DEEP feature sets) did not show significant improvements over QUEST-17 models for the majority of the cases in all language pairs. Finally, the use of different kernels and GP did not improve over the SVR counterpart (with only one kernel) for all the language pairs. The best QE systems predicting the different quality labels (BLEU, TER and METEOR) were very similar for each language pair, in terms of Pearson's *r* correlation scores. For example, for EN-DE, the best systems were built with ALL features and GP for all labels. A similar behaviour was found by the systems for EN-ES, EN-FR and ES-EN. For DE-EN and FR-EN there are some small differences, but the best systems are always very similar among the different labels. There is also no correlation between the performance of the MT systems (Table 5.5) and the results for QE. As we mentioned during the presentation of the results per language pair, the overall quality of the MT systems did not lead to better performance on the QE models.

5.4 Experiments with HTER: LIG corpus

In this section, we experiment with the HTER metric as a measure of document quality and compare its results with BLEU, TER and METEOR.

Corpus The corpus used is the LIG corpus (Potet et al., 2012) presented in Chapter 4. For the experiments of this section, we use an extended version of the corpus with 361 FR-EN documents. We use 70% of the documents for training the QE systems and 30% for test.

Features For this corpus, there were no pseudo-references available and, as a consequence, CONSENSUS features were not used. Although pseudo-references could be generated by using off-the-shelf online MT systems (such as Google Translate), we avoid this approach for this corpus because it is expected that this data is already in the database of such systems, which would give us a reference translation and not a machine translation. DEEP features are also not used since English is the target language. Therefore, the feature sets used are: QUEST-17, QUEST-ALL, QUEST-17+SHALLOW, QUEST-ALL+SHALLOW, QUEST-17-WE and ALL.

Quality labels The automatic metrics selected for quality labelling and prediction are HTER, BLEU, TER and METEOR. The Asiya toolkit was used to calculate all metrics.

Method SVR and GP are used to generate the QE models.

Results Table 5.12 shows the results for the LIG corpus. Apart from two models predicting HTER with GP and QUEST-17 and GP and QUEST-17+WE, all Pearson's r correlation scores are significant. The highest correlation scores for predicting BLEU were achieved by the ALL feature set and SVR model. For predicting TER, the two models with the ALL feature set showed the highest correlation scores. Finally, for predicting METEOR the highest correlation score was achieved using ALL features and a GP model. The prediction of HTER caused several systems to perform very similarly. Finally, the use of different kernels and GP models (for QUEST-17+WE and QUEST-17+ALL) did not outperform the SVR counterparts with only one kernel for all feature sets.

Figure 5.10 shows the performance gains for experiments with the LIG corpus. Best results differ among the features, but no system achieved performance gains higher than 20% (gains are even smaller for BLEU and TER). SHALLOW features showed performance gains comparable to those of the best models.

	BLEU		TER		METEOR		HTER	
	SVR	GP	SVR	GP	SVR	GP	SVR	GP
QUEST-17	0.239	0.211	0.211	0.229	0.468	0.471	**0.386**	0.174*
QUEST-ALL	0.297	0.363	0.307	0.321	0.506	0.523	0.357	0.215
QUEST-17+SHALLOW	0.336	0.306	0.308	0.300	0.522	0.516	**0.369**	0.259
QUEST-ALL+SHALLOW	0.297	0.383	0.314	0.346	0.510	0.539	**0.372**	0.272
QUEST-17+WE	0.423	0.389	0.401	0.372	0.573	0.565	**0.366**	0.149*
ALL	**0.436**	0.391	**0.411**	0.402	0.583	**0.597**	**0.377**	0.249
RANDOM	0.009*		0.100*		0.011*		0.071*	

Table 5.12 Results for LIG in terms of Pearson's r correlation. * indicates results that did not show significant Pearson's r correlation with p-value < 0.05. The best systems are highlighted in bold (William's significance test with p-value < 0.05).

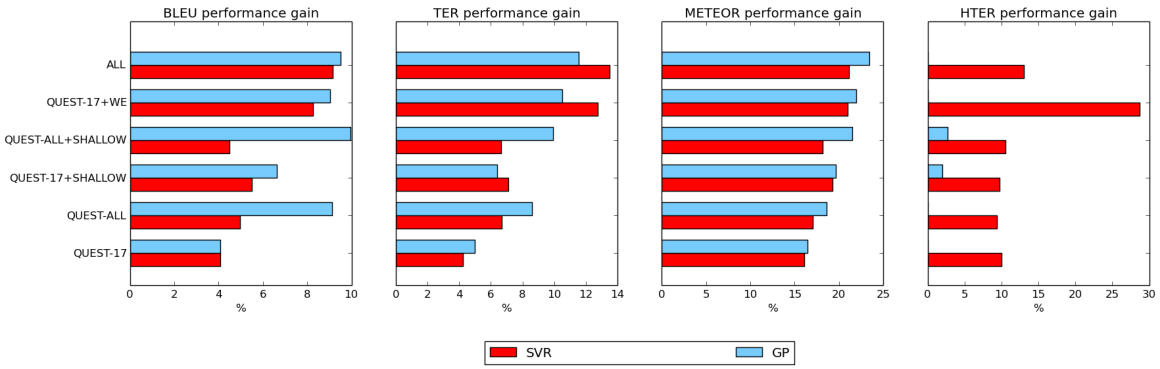

Fig. 5.10 Performance gains in terms of MAE of the QE models for LIG documents.

5.5 Problems with Automatic Metrics as Labels for Document-level QE

In the previous sections we presented several experiments for document-level QE, varying feature types, ML approaches and datasets. However, the quality labels used were traditional automatic evaluation metrics (BLEU, TER and METEOR) or standard human-targeted metrics (HTER). Although some of our models showed moderate or strong correlation scores with true labels according to Pearson's r and some models produced considerable performance gains, we need to examine the data closely in order to understand such results. It's known that BLEU-style metrics fail to provide a fair evaluation of MT system outputs that differ from the human reference (Callison-Burch, Osborne, and Koehn, 2006). In addition, such metrics are designed for system-level evaluation (i.e. for comparing different

MT systems on the same data) and, therefore, they also fail in scoring single documents, according to document-level issues.

One result that supports our hypothesis is the fact that discourse-aware features did not outperform the baseline. In fact, for some language pairs in the WMT experiments, document-aware features could not even achieve the baseline score. Discourse information is expected to be impacted by machine translation and, therefore, discourse features should show promising results.[10]

We conducted an analysis over all datasets and found that the majority of BLEU, TER, METEOR and even HTER values are concentrated on a small range of the data. Therefore, all documents are treated as having similar quality. This assumption is also supported by the fact that all MAE_{mean} were below 0.1 in our experiments. This means that using a predictor that assigns the mean value of the training set as the predicted value for all entries in the test set already solves the prediction problem with considerably low error. Finally, the reason why the models trained in the previous section show good performance for predicting the automatic metrics (in terms of Pearson's r correlation) may be because the predictions, as well as the true labels, are clustered around a small area.

The data was analysed in terms of statistical dispersion and central tendency, by computing mean, standard deviation (STDEV), median and interquartile range (IQR). In order to support the results, we also present the maximum and minimum values as well as the values of the first (Q_1) and third (Q_3) quartiles.

Mean is a central tendency metric that is calculated by totaling all elements in the sample and dividing it by the total number of elements (Equation 5.3 - where N is the number of elements in the sample). Since it is a simple sum of all values, it is not robust to outliers, meaning that outliers may have a significant impact on the mean. STDEV is the square root of the variance, a statistical dispersion metric that measures how far the elements are from the mean. Since this metric is based on the mean value, it is also not robust to outliers (Equation 5.4).

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{5.3}$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{5.4}$$

[10] Assuming our discourse features are accurate representations of discourse information.

Median is another central tendency metric and is defined by the central element in the sample (Equation 5.5). The sample is put in ascending order before computing the median.

$$median = \frac{n-1}{2} th \text{ element} \tag{5.5}$$

When the sample has an even number of elements, the median is defined as the average of the two central values. This metric is less sensitive to outliers and, therefore, it is consider robust.

IQR is a statistical dispersion metric that consists of the difference between the third and first quartiles (Equation 5.6). Quartiles are the three points able to divide the dataset in equal groups, each containing a quarter of the data. The first quartile (Q_1) is the point that divides the 25% smallest values from the rest of the data. The second quartile corresponds to the point that divides the data in two equal sized groups (50%) and it corresponds to the median value. Finally, all data points with a value higher than the third quartile (Q_3) correspond to 25% of the data with the highest values. IQR shows the interval where 50% of the data appear.

$$IQR = Q_3 - Q_1 \tag{5.6}$$

FAPESP Table 5.13 shows the statistics of the three scenarios in the FAPESP corpus, for all quality labels (BLEU, TER and METEOR). All automatic evaluation metrics show reasonable STDEV in relation to the mean (the ratio between STDEV and mean is around 0.200). However, the results given by IQR show much lower dispersion. IQR showed distances smaller than 0.130 for MOSES and SYSTRAN and smaller than 0.150 for MIXED, meaning that 50% of the data appear in a small interval. For this dataset, it is important to note that by mixing different MT systems in the same dataset (MIXED), the dispersion metrics presented higher values. However, they are not far from the cases with only one MT system. Perhaps BLEU-style metrics are not evaluating the documents in the way needed for them to distinguish between different documents.

WMT Table 5.14 shows the statistical metrics for all language pairs of the WMT dataset. For all language pairs, STDEV shows considerable data variation. However, IQR values were small. For WMT EN-DE, IQR were 0.079, 0.153 and 0.094 for BLEU, TER and METEOR respectively. This means that 50% of the data points are located on a small interval. For EN-ES, the IQR values are slightly higher than for EN-DE in each metric, with the highest

MOSES								
	MIN	MAX	MEAN	STDEV	Q_1	MEDIAN	Q_3	IQR
BLEU	0.005	0.589	0.361	**0.075**	0.316	0.364	0.416	**0.100**
TER	0.265	1.000	0.497	**0.098**	0.427	0.481	0.549	**0.123**
METEOR	0.060	0.726	0.541	**0.074**	0.499	0.548	0.592	**0.093**
SYSTRAN								
	MIN	MAX	MEAN	STDEV	Q_1	MEDIAN	Q_3	IQR
BLEU	0.007	0.831	0.275	**0.058**	0.237	0.279	0.313	**0.076**
TER	0.107	1.000	0.557	**0.090**	0.496	0.539	0.608	**0.113**
METEOR	0.054	0.891	0.480	**0.066**	0.440	0.488	0.524	**0.084**
MIXED								
	MIN	MAX	MEAN	STDEV	Q_1	MEDIAN	Q_3	IQR
BLEU	0.004	0.687	0.364	**0.104**	0.287	0.353	0.441	**0.153**
TER	0.201	1.000	0.483	**0.118**	0.396	0.480	0.553	**0.157**
METEOR	0.055	0.806	0.552	**0.118**	0.488	0.546	0.624	**0.136**

Table 5.13 Statistic dispersion and central tendency metrics for the FAPESP dataset.

being 0.201 for TER. However, 0.201 is still a low variation. The other language pairs present a similar behaviour for IQR. The highest IQR value overall is of 0.220 for TER for EN-FR. In general, TER obtains the highest values of IQR and STDEV.

LIG For the LIG dataset, the difference between the minimum and maximum values are considerably smaller for BLEU and METEOR than for TER and HTER. On the other hand, the results are similar with all automatic metrics in terms of IQR, achieving a low dispersion. The highest scores where for TER with 0.138 of IQR, although this is still a sign of low variation.

We also analyse the distributions of the training and test sets. For all corpora, language pairs and quality labels, the training and test data follow the same distribution, according to the Kolmogorov-Smirnov test.[11] In addition, the data distribution for all scenarios does not follow the normal distribution (also according to the Kolmogorov-Smirnov test). Therefore, since there are no differences between the distributions of training and test sets and the data distribution does not follow the normal distribution for all cases, there is no evidence of the impact of these factors into the QE approaches.

In order to further validate our assumptions regarding the variation in the data, we also plot the distribution of the true values and predicted values for some scenarios. The predicted values are from the best systems presented in the previous sections. When two systems have the same performance (for example, HTER prediction in the LIG dataset) one of the systems were selected randomly. In this section, we only present the best results for the FAPESP

[11]The null hypothesis that the two samples are drawn from the same distribution was not rejected with p-value > 0.05.

EN-DE								
	MIN	MAX	MEAN	STDEV	Q_1	MEDIAN	Q_3	IQR
BLEU	0.000	0.325	0.130	**0.058**	0.088	0.127	0.168	**0.079**
TER	0.506	1.000	0.830	**0.105**	0.758	0.835	0.911	**0.153**
METEOR	0.000	0.542	0.322	**0.081**	0.278	0.325	0.372	**0.094**
EN-ES								
	MIN	MAX	MEAN	STDEV	Q_1	MEDIAN	Q_3	IQR
BLEU	0.056	0.681	0.245	**0.102**	0.167	0.234	0.313	**0.146**
TER	0.215	1.000	0.649	**0.138**	0.548	0.648	0.749	**0.201**
METEOR	0.234	0.819	0.480	**0.103**	0.408	0.474	0.549	**0.142**
EN-FR								
	MIN	MAX	MEAN	STDEV	Q_1	MEDIAN	Q_3	IQR
BLEU	0.000	1.000	0.232	**0.124**	0.151	0.216	0.295	**0.144**
TER	0.000	1.000	0.689	**0.163**	0.578	0.694	0.799	**0.220**
METEOR	0.000	1.000	0.433	**0.131**	0.354	0.428	0.512	**0.158**
DE-EN								
	MIN	MAX	MEAN	STDEV	Q_1	MEDIAN	Q_3	IQR
BLEU	0.000	1.000	0.186	**0.096**	0.130	0.176	0.227	**0.096**
TER	0.000	1.000	0.606	**0.124**	0.529	0.604	0.678	**0.149**
METEOR	0.000	1.000	0.281	**0.071**	0.248	0.278	0.308	**0.059**
ES-EN								
	MIN	MAX	MEAN	STDEV	Q_1	MEDIAN	Q_3	IQR
BLEU	0.040	1.000	0.244	**0.134**	0.155	0.226	0.303	**0.148**
TER	0.000	0.940	0.523	**0.151**	0.424	0.523	0.625	**0.201**
METEOR	0.180	1.000	0.321	**0.093**	0.266	0.311	0.359	**0.092**
FR-EN								
	MIN	MAX	MEAN	STDEV	Q_1	MEDIAN	Q_3	IQR
BLEU	0.000	1.000	0.230	**0.105**	0.158	0.216	0.297	**0.139**
TER	0.000	1.000	0.538	**0.137**	0.445	0.533	0.628	**0.183**
METEOR	0.000	1.000	0.309	**0.070**	0.272	0.304	0.346	**0.074**

Table 5.14 Statistic dispersion and central tendency metrics for the WMT dataset.

	MIN	MAX	MEAN	STDEV	Q_1	MEDIAN	Q_3	IQR
BLEU	0.062	0.590	0.251	**0.078**	0.195	0.253	0.298	**0.104**
TER	0.159	0.820	0.505	**0.103**	0.434	0.499	0.573	**0.138**
METEOR	0.184	0.459	0.327	**0.050**	0.289	0.329	0.364	**0.075**
HTER	0.111	1.000	0.244	**0.078**	0.204	0.234	0.275	**0.071**

Table 5.15 Statistic dispersion and central tendency metrics for the LIG dataset.

dataset in the MIXED scenario, for WMT EN-DE and DE-EN and for LIG (similar plots are obtained for the other cases).

Figures 5.11, 5.12, 5.13 and 5.14 show the distributions for MIXED, WMT EN-DE, WMT DE-EN and LIG, respectively. The majority of the distributions show that, although the predicted values follow the distribution of the true values (and, consequently, show considerable correlation), the data distribution is located in a very small portion of the data spectrum. Therefore, a classifier that always predicts a number in this small interval would probably obtain good results. The exceptions are BLEU, TER and METEOR for

MIXED in Figure 5.11 and TER for WMT DE-EN (Figure 5.13), which show a considerably wider spread. There is no clear correlation between the performance of the QE systems and the variation of the data. The FAPESP case showed the best QE systems in terms of Pearson's *r* correlation, whilst also presented the highest variation. On the other hand, the performance of the best QE systems for the WMT DE-EN data predicting TER is not far from the performance of the best systems predicting BLEU and METEOR, even though the TER distribution has more variation than the distributions for BLEU and METEOR. Therefore, since the data variation of the automatic evaluation metrics is considerably different in different datasets, we conclude that such metrics are problematic. Based on these observations, it is unclear whether or not the automatic evaluation metrics explored herein can be reliably used as quality labels for predicting document-level quality.

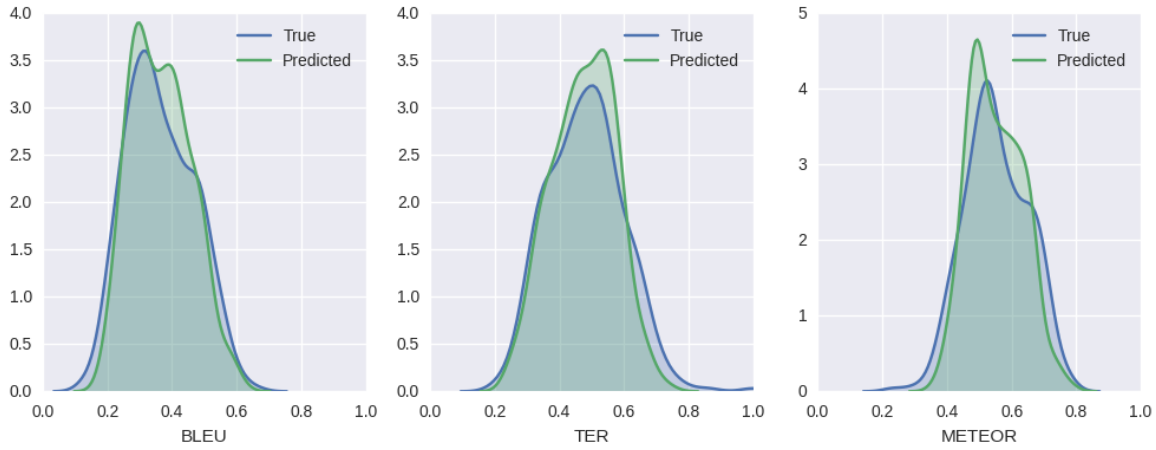

Fig. 5.11 Data distribution of true and predicted values of the best systems predicting BLEU, TER and METEOR for MIXED scenario in the FAPESP dataset.

5.6 Discussion

In this chapter, we presented several experiments with different datasets for document-level QE, exploring the features introduced in Chapter 4. The quality labels used were BLEU, TER and METEOR for all datasets except LIG, for which we also including the HTER metric. In general, for all quality labels, the addition of discourse-aware features (feature sets DEEP and SHALLOW) did not improve over the baseline systems built with only the QUEST-17 feature set.

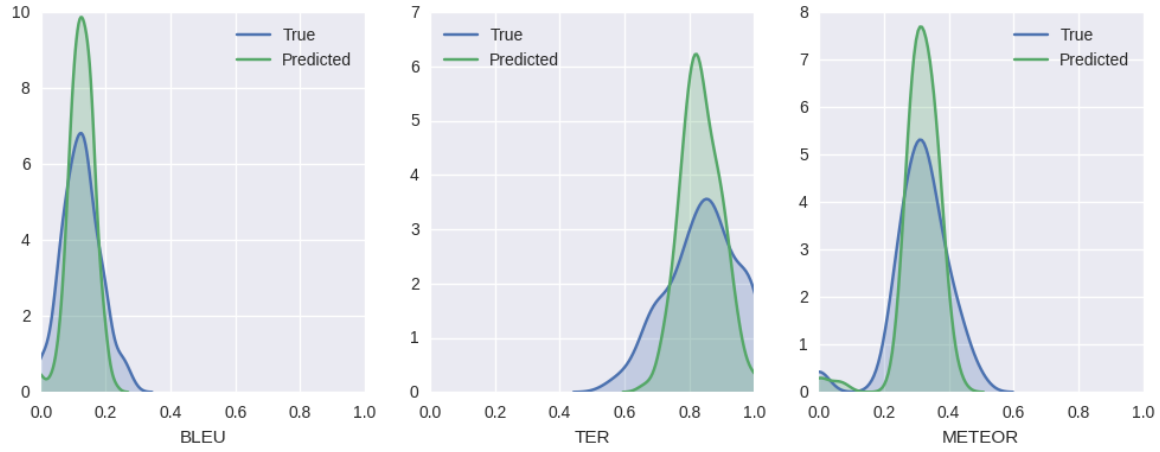

Fig. 5.12 Data distribution of true and predicted values of the best systems predicting BLEU, TER and METEOR for EN-DE in the WMT dataset.

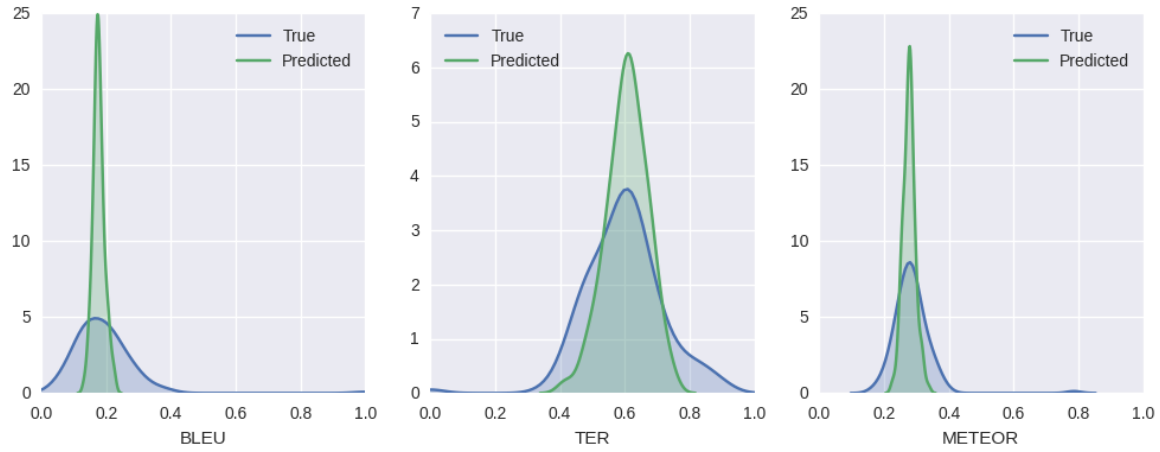

Fig. 5.13 Data distribution of true and predicted values of the best systems predicting BLEU, TER and METEOR for DE-EN in the WMT dataset.

For the FAPESP dataset (Section 5.2), the use of PSEUDO or CONSENSUS feature sets achieved the best results, which are similar to the current state-of-the-art results for document-level QE presented in Chapter 2. However, as mentioned in Chapter 4, pseudo-reference-based features are not available in realistic scenarios and therefore we consider the merit of these results not to be impressive as those of other feature sets. SVR models were consistently better than GP models in this dataset, although the same cannot be observed in the results obtained over the WMT dataset.

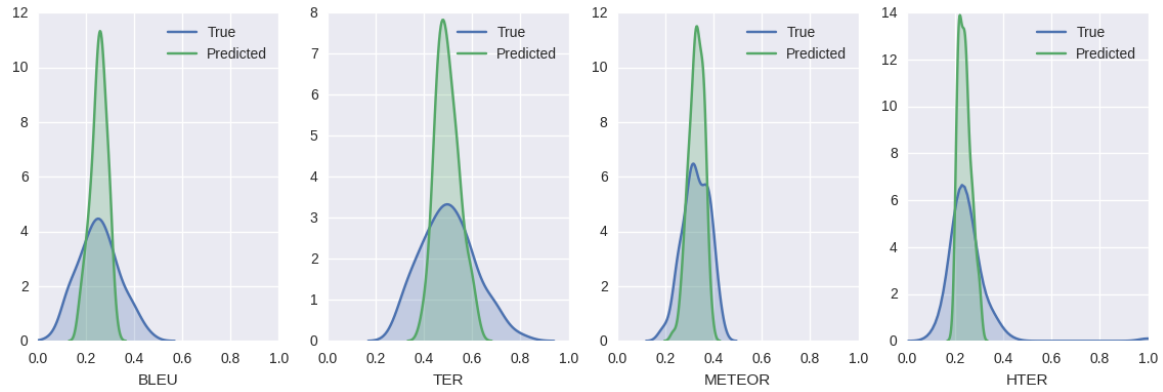

Fig. 5.14 Data distribution of true and predicted values of the best systems predicting BLEU, TER, METEOR and HTER for LIG dataset.

For the WMT dataset, the combination of all features (in feature set ALL) led to the best results in the majority of the cases (Section 5.3). The best ML technique varied across the different language pairs and, therefore, we cannot conclude on what are the best techniques for the task and recommend that both algorithms are tested for each dataset. SVR models are, in general, faster to train and test than GP models, although the difference in time (or performance) was not very salient.

For the LIG dataset, models built with the ALL feature set showed the best results for BLEU, TER and METEOR (Section 5.4). The best ML technique varied among the different labels: SVR was the best for BLEU, GP was the best for METEOR and both had similar performance for TER. For HTER almost all models built with SVR did not differ significantly (the only exception was the model built with QUEST-ALL features). The variance in the results over different datasets do not allow us to generalise which configuration is the most effective for the task. However, such differences can be a result of the use of automatic metrics as quality labels. We believe that such metrics are not reliable for the task and this could explain why there is no consensus across the different datasets.

Finally, in Section 5.5 we show the variation in the data distribution of all automatic metrics in all datasets. The scores obtained for IQR are small across all scenarios evaluated, meaning that a large part of the data points are located in a small interval. STDEV showed considerably high variation, which together with the IQR results highlight outliers in the data. However, the distribution of some datasets showed a considerably wide range where the scores are distributed (such as the MIXED in FAPESP dataset - Figure 5.11). On the other hand, for other datasets, this range of the data is very small (for example, HTER for the LIG dataset - Figure 5.14). Therefore, it is not clear how reliable are such automatic evaluation

metrics for the assessment of documents, since their distributions varied a lot across different datasets. Moreover, it would be expected that QE models showed better results in terms of Pearson's *r* correlation and performance gains for datasets with low variance, since the prediction task would be easier for the models. However, the dataset with best correlation scores and performance gains is the one with the highest data variation (FAPESP). In addition, the performance of the QE systems varied a lot for different datasets with similar low variance in the data distribution (for examples, BLEU, TER and METEOR for EN-DE and DE-EN in the WMT dataset). Therefore, there might be factors other than data variation that impact the performance of the QE models. Since the features are compatible across different datasets and language pairs, our hypothesis is that the labels are not ideal for the task.

In the next chapter we propose and experiment with new task-based document-aware quality labels. Models are trained with configuration similar to those of the models built in this chapter and the feature sets used are the same (apart from when we approach QE as a classification task). Whenever it is possible we also compare the models built for predicting our new labels with models built for predicting BLEU-style metrics.

Chapter 6

New Quality Labels for Document-level QE

As discussed in Chapter 2, defining "quality" is vital for the success of the evaluation of NLP tasks.[1] In MT, the notion of quality can vary considerably depending on the audience and the purposes of the translation. Although traditional evaluation metrics (e.g. BLEU) can be considered useful for system comparisons and system evaluations, they appear not to be reliable for document-level assessment (Chapter 5). Such metrics operate at either sentence or corpus level and they do not take into account document-wide problems (such as discourse).

The experiments presented in Chapter 5 show that there is a lack of reliable quality labels for document-level QE, which could be used to train QE models or to compare automatic metrics against. In this Chapter we introduce novel quality labels specifically designed for document-level QE, which we believe to be more reliable for the QE task. Two different types of quality labels are presented: one for dissemination purposes and another for assimilation.

Section 6.1 shows the first experiment done towards new quality labels for document-level QE. We experiment with direct human assessments of cohesion and coherence and introduce a new method of two-stage post-editing.

Section 6.2 presents our large-scale experiments with the two-stage post-editing method, aiming to devise quality labels for dissemination purposes. Two new labels are devised by combining edit distance scores of the post-editing tasks and compared with BLEU, TER and METEOR as quality labels.

[1] Parts of this chapter were previously published in peer-reviewed conferences: Scarton et al. (2015), Scarton, Tan, and Specia (2015), Scarton and Specia (2016) and Scarton et al. (2016).

In Section 6.3 we introduce new quality labels derived from reading comprehension tests for assimilation purposes. The marks for the reading comprehension tests are converted into quality scores in two ways: (i) for open questions we use continuous marking scores and (ii) for multiple choice questions we use discrete marking scores. Therefore, the problem is addressed either as regression (for continuous scores) or as classification (for discrete scores).

6.1 Preliminary Experiments

The first experiments done towards devising labels for document-level QE were conducted at paragraph level. We considered paragraphs as small documents since they usually address a single topic and the sentences are connected coherently. In other words, paragraphs are long enough to encompass certain document-level information and short enough to make experiments and analyses feasible. We ran two experiments at paragraph level: human assessment of cohesion and coherence (called hereafter SUBJ) and two-stage post-editing (called here after PE1, first post-editing, and PE2, second post-editing).

6.1.1 Experimental Settings

Data

The datasets were extracted from the test set of the EN-DE WMT13 MT shared task. EN-DE was chosen given the availability of in-house annotators for this language pair. Outputs of the **UEDIN** SMT system were chosen as this was the highest performing participating system for this language pair (Bojar et al., 2013). For the SUBJ experiment, 102 paragraphs were randomly selected from the FULL CORPUS (Table 6.1).

For PE1 and PE2, only source (English) paragraphs with 3-8 sentences were selected (filter S-NUMBER) to ensure that there was enough information beyond sentence-level to be evaluated and make the task feasible for the annotators. These paragraphs were further filtered to discard those without cohesive devices. Cohesive devices are linguistic units that play a role in establishing cohesion between clauses, sentences or paragraphs (Halliday and Hasan, 1976). Pronouns and discourse connectives are examples of such devices. A list of pronouns and the connectives from Pitler and Nenkova (2009) was used. Finally, paragraphs were ranked according to the number of cohesive devices they contain and the top 200 paragraphs were selected (filter C-DEV). Table 6.1 shows the statistics of the initial corpus and the resulting selection after each filter.

	Number of Paragraphs	Number of Cohesive devices
FULL CORPUS	1,215	6,488
S-NUMBER	394	3,329
C-DEV	200	2,338

Table 6.1 WMT paragraph-level corpus statistics.

For the PE1 experiment, the paragraphs in C-DEV were randomised. Then, sets containing seven paragraphs each were created. For each set, the sentences of its paragraphs were also randomised in order to prevent annotators from having access to wider context when post-editing. The guidelines made it clear to annotators that the sentences they were given were not related, not necessarily part of the same document, and that therefore they should not try to find any relationships among them. For PE2, sentences that had already been post-edited were put together in their original paragraphs and presented to the annotators as a complete paragraph. In PE2, the annotators post-edited the same sentences that they had already post-edited in PE1.

Annotators

The annotators for both experiments were students of "Translation Studies" courses (TS) in Saarland University, Saarbrücken, Germany. All students were familiar with concepts of MT and with post-editing tools. They were divided into two groups: (i) 25 *Undergraduate students (B.A.)*, who are native speakers of German; and (ii) 29 *Master students (M.A.)*, the majority of whom are native speakers of German. Non-native speakers had at least seven years of German language studies. B.A. and M.A. students have on average 10 years of English language studies. Only the B.A. group did the SUBJ experiment. PE1 and PE2 were done by all groups.

PE1 and PE2 were done using three CAT tools: PET (Aziz, Sousa, and Specia, 2012), Matecat (Federico et al., 2014) and memoQ.[2] These tools operate in very similar ways in terms of their post-editing functionalities, and therefore the use of multiple tools was only meant to make the experiment more interesting for students and did not affect the results. SUBJ was done without the help of tools.

For PE1, the only guideline provided was that the annotators should perform corrections in the MT output without focusing on style. They should only focus in making the text fluent and coherent with the source document. For PE2, we asked the annotators to make

[2]https://www.memoq.com/

any remaining corrections that were necessary for keeping the coherence with the source document.

6.1.2 Human Assessments: Cohesion and Coherence

Our first attempt to access quality beyond sentence level was to explicitly guide annotators to consider discourse, where the notion of "discourse" covers various linguistic phenomena observed across discourse units. Discourse units can be clauses (intra-sentence), sentences or paragraphs. Consequently, the SUBJ experiment consists in assessing the quality of paragraphs in terms of cohesion and coherence. We define cohesion as the linguistic marks (cohesive devices) that connect clauses, sentences or paragraphs together; coherence captures whether clauses, sentences or paragraphs are connected in a logical way, i.e. whether they make sense together (Stede, 2011). In order to assess these two phenomena, we propose a 4-point scale that is similar to the human assessment scale for fluency and adequacy. For coherence:

1. Completely coherent;
2. Mostly coherent;
3. Little coherent;
4. Incoherent.

For cohesion:

1. Flawless;
2. Good;
3. Disfluent;
4. Incomprehensible.

Six sets with 17 paragraphs each were randomly selected from the 102 paragraphs and given to the 25 annotators from the B.A. group (each annotator evaluated one set). The task was to assess the paragraphs in terms of cohesion and coherence, using the scale given. The annotators could also rely on the source paragraphs (the guidelines of this experiment are presented in Appendix B)

The agreement for the task in terms of Fleiss-Kappa[3] and Spearman's ρ rank correlation are given in Tables 6.2 and 6.3, respectively. The number of annotators per set is different because some of them did not complete the task.

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6
Annotators	3	3	4	7	6	2
Coherence	0.05	0.00	0.06	0.06	0.06	0.35
Cohesion	0.13	0.37	0.14	0.05	0.05	0.13

Table 6.2 Fleiss inter-annotator agreement for the SUBJ task.

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6
Annotators	3	3	4	7	6	2
Coherence	0.07	0.05	0.16	0.16	0.28	0.58
Cohesion	0.38	0.43	0.28	0.09	0.38	0.12

Table 6.3 Spearman's ρ rank correlation for the SUBJ task.

The Fleiss-Kappa agreement is low for all the sets, but cohesion on Set 3 is an exception (although the correlation is still below 0.5). A low agreement in terms of Spearman's ρ rank correlation was found for coherence (ranging from 0.05 to 0.28, having 0.58 as an outlier). For cohesion, half of the sets show moderate correlation whilst the other half show weak or no correlation for Spearman's ρ (using the same classification presented in Table 5.1 from Chapter 5 for Pearson's r). These can be sign that cohesion is easier to be assessed than coherence. However, these concepts are naturally very abstract, even for humans, offering substantial room for subjective interpretations. In addition, the existence of (often many) errors in the MT output can hinder the understanding of the text altogether, rendering judgements on any specific quality dimension difficult to make.

6.1.3 Two-stage Post-editing

In order to overcome the issues of direct human evaluation, we propose a new method that we call two-stage post-editing (Scarton et al., 2015). Such a method is based on a human-targeted task-based approach that indirectly evaluates machine translated documents by using human post-editing.

[3]This metric is an extension of the Kappa metric allowing agreement calculations over more than two annotators.

Similarly to HTER (Snover et al., 2006), the main idea here is to have humans post-edited machine translations and use the edit distance between the original MT output and its post-edited version as a proxy to quality. However, different from HTER, we aim to isolate errors that are local to a sentence from errors that involve discourse-wide information and incorporate these two components into the error metric. Given a collection with multiple documents, all sentences from all documents are put together and randomly shuffled. As it was described in Section 6.1.1, in the first stage of the method (PE1), these randomised sentences are given to the translators for annotation. The only information provided is the source sentence and the machine translated sentence to be corrected. In the second stage (PE2), the post-edited sentences are put together, forming the original document, and the same translators are asked to correct any remaining errors. We explore the assumption that the latter errors can only be corrected with document-wide context. The final document is fully post-edited such that it could be used for dissemination purposes.

We perform a first exercise of this method at paragraph level, with C-DEV data presented in Table 6.1. Using HTER, we measured the edit distance between the post-edited versions with and without context. The hypothesis we explore is that the differences between the two versions are likely to be corrections that could only be performed with information beyond sentence level. A total of 112 paragraphs were evaluated in 16 different sets, but only sets where more than two annotators completed the task are presented here (SET1, SET2, SET7, SET9, SET14 and SET15) (see Table 6.4).[4]

Task Agreement

Table 6.4 shows the agreement for the PE1 and PE2 tasks using Spearman's ρ rank correlation. The correlations were calculated by comparing the HTER values of PE1 against MT and PE2 against PE1 for each set. In Table 6.4 we also present the averaged HTER for each set, among all annotators. The values for HTER among annotators in PE2 against PE1 were averaged in order to provide a better visualisation of changes made in the paragraphs from PE1 to PE2.

The HTER values of PE1 against PE2 are low, as expected, since the changes from PE1 to PE2 are only expected to reflect discourse related issues. In other words, no major changes were expected during the PE2 task, although some sets show a considerable amount of edits (for example, SET9). The Spearman's ρ correlation scores for HTER between PE1 and MT varies from 0.22 to 0.56, whereas the correlation in HTER between PE1 and PE2 varies between -0.14 and 0.39. The negative figures mean that the annotators strongly disagreed

[4]Sets with only two annotators are difficult to interpret.

	SET1	SET2	SET5	SET6	SET9	SET10	SET14	SET15	SET16
Annotators	3	3	3	4	4	3	3	3	3
PE1 x MT - averaged HTER	0.63	0.57	0.22	0.32	0.28	0.18	0.30	0.24	0.18
PE1 x PE2 - averaged HTER	0.05	0.07	0.05	0.03	0.10	0.06	0.09	0.07	0.05
PE1 x MT - Spearman	0.52	0.50	0.52	0.56	0.37	0.41	0.71	0.22	0.46
PE2 x PE1 - Spearman	0.38	0.39	−0.03	−0.14	0.25	0.15	0.14	0.18	−0.02

Table 6.4 Averaged HTER values and Spearman's ρ rank correlation for PE1 against MT and PE1 against PE2.

regarding the changes made from PE1 to PE2. This can be related to stylistic choices made by the annotators (see Section 6.1.3).

Issues Beyond Sentence Level

Figure 6.1 shows the results for individual paragraphs in all sets. The majority of the paragraphs were edited in the second round of post-editions. This clearly indicates that information beyond sentence-level can be helpful to further improve the output of MT systems. Between 0% and 19% of the words have changed from PE1 to PE2 depending on the paragraph and the annotators (in average, 7% of the words were edited).

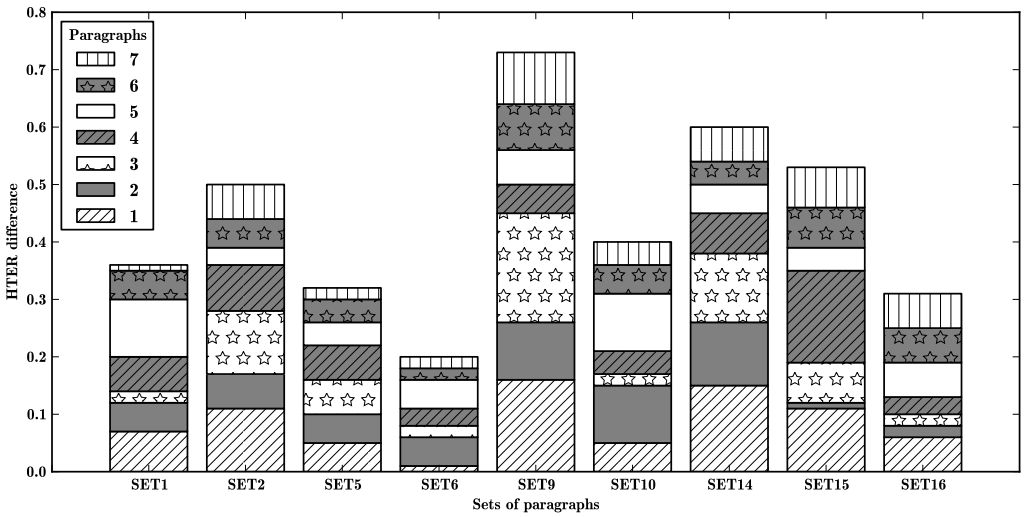

Fig. 6.1 HTER between PE1 and PE2 for each of the seven paragraphs in each set.

An example of changes from PE1 to PE2 related to discourse phenomena is shown in Table 6.5. In this example, two changes were performed. The first is related to the substitution of the sentence *"Das ist falsch"* - literal translation of *"This is wrong"* - by *"Das ist nicht gut"*, which fits better into the context, because it gives the sense of *"This is not good"*.

The other change makes the information more explicit. The annotator decided to change from *"Hier ist diese Schicht ist dünn"* - literal translation of *"Here, this layer is thin"* - to *"Hier ist die Anzahl solcher Menschen gering"*, a translation that better fits the context of the paragraph: *"Here, the number of such people is low"*.

PE1: - St. Petersburg bietet nicht viel kulturelles Angebot, Moskau hat viel mehr Kultur, es hat eine Grundlage. Es ist schwer für die Kunst, sich in unserem Umfeld durchzusetzen. Wir brauchen das kulturelle Fundament, aber wir haben jetzt mehr Schriftsteller als Leser. **Das ist falsch**. In Europa gibt es viele neugierige Menschen, die auf Kunstausstellungen, Konzerte gehen. **Hier ist diese Schicht ist dünn**.
PE2: - St. Petersburg bietet nicht viel kulturelles Angebot, Moskau hat viel mehr Kultur, es hat eine Grundlage. Es ist schwer für die Kunst, sich in unserem Umfeld durchzusetzen. Wir brauchen das kulturelle Fundament, aber wir haben jetzt mehr Schriftsteller als Leser. **Das ist nicht gut**. In Europa gibt es viele neugierige Menschen, die auf Kunstausstellungen, Konzerte gehen. **Hier ist die Anzahl solcher Menschen gering.**
SRC: - St. Petersburg is not a cultural capital, Moscow has much more culture, there is bedrock there. It's hard for art to grow on our rocks. We need cultural bedrock, but we now have more writers than readers. **This is wrong**. In Europe, there are many curious people, who go to art exhibits, concerts. **Here, this layer is thin.**

Table 6.5 Example of changes from PE1 to PE2.

Manual Analysis

In order to better understand the changes made by the annotators from PE1 to PE2 and also better explain the negative values in Table 6.4, we manually inspected the post-edited data. This analysis was done by senior translators who were not involved in the actual post-editing experiments. They counted modifications performed and categorised them into three classes:

Discourse/context changes: changes related to discourse phenomena, which could only be made by having the entire paragraph.

Stylistic changes: changes related to translator's stylistic or preferential choices. These changes can be associated with the paragraph context, although they are not strictly necessary under our post-editing guidelines.

Other changes: changes that could have been made without the paragraph context (PE1), but were only performed during PE2.

The results are shown in Table 6.6. Although annotators were asked not to make unnecessary changes (stylistic), some of them made changes of this type (especially annotators 2 and 3 from sets 5 and 6, respectively). These sets are also the ones that show negative values in Table 6.4. Since stylistic changes do not follow a pattern and are related to the background and preferences of the translator, the high number of this type of change for these sets may be the reason for the negative correlation figures. In the case of SET6, annotator 2 also performed several changes classified as "other changes". This may have also led to negative correlation values. However, the reasons behind the negative values in SET16 could include other phenomena, since overall the variation in the changes performed is low.

	SET1			SET2			SET5			SET6				SET9				SET10			SET14			SET15			SET16		
Annotators	1	2	3	1	2	3	1	2	3	1	2	3	4	1	2	3	4	1	2	3	1	2	3	1	2	3	1	2	3
Discourse/context	2	3	1	0	6	2	2	1	0	2	2	0	0	1	7	1	0	4	0	0	1	0	1	2	1	2	0	1	1
Stylistic	2	0	1	1	0	1	3	11	0	0	3	9	3	5	10	1	3	1	2	2	6	0	0	3	3	2	2	1	3
Other	1	2	4	0	2	2	2	2	6	0	6	0	1	2	0	4	2	1	0	2	2	0	1	1	2	1	1	1	0
Total errors	5	5	6	1	8	5	7	14	6	2	11	9	4	8	17	6	5	6	2	4	9	0	2	6	6	5	3	3	4

Table 6.6 Counts on types of changes made from PE1 to PE2.

In the next section (Section 6.2) we show our large scale experiments with the two-stage post-editing method. We also rely on professional post-editors, with more experience than the annotators used in the preliminary study presented here. Moreover, we controlled the post-editions by monitoring the annotators during the annotation task.

6.2 Dissemination: Two-stage Post-editing

The two-stage post-editing method used in this section is an extension of the experiments done in Section 6.1.3 for a different language pair (EN-ES instead of EN-DE), full documents (instead of paragraphs), more data and relying on professional translators. For our new experiment, 208 documents were selected from the EN-ES WMT translation shared task corpora (2008-2013 editions), with an average of 302.88 words per document and 15.76 sentences per document. These documents are a subset of the WMT EN-ES dataset used in Chapter 5, therefore, the machine translations come from different MT systems. The selection was made such that it ensures that shortest documents in the collection are featured first, for practical reasons (time and cost of post-editing) as well as so that more documents could be annotated.

The documents were post-edited by two professional translators (half the documents each), who were hired to work full time for two weeks in the project. The post-editing guidelines were similar to the ones presented in Section 6.1.1, although the environment

in this experiment was more controlled. Firstly, the post-editors performed a training task, in order to solve any questions about the two-stage post-editing method and to make them familiar with the post-editing tool (PET). Secondly, the post-editors worked on site with us and they gave us daily reports on the task development.[5]

Deriving Quality Labels

We evaluate four variants of edit-distance-based quality labels, called hereafter: DISS-HTER, DISS-LC-P, DISS-LC-M and DISS-LC-W2. For **DISS-HTER**, we used the standard HTER scores between PE2 and MT. In this case, the quality label gives the same importance to all edits that are made in the document, i.e, word, sentence and document-level problems are given the same weight. Although DISS-HTER is calculated simply as the HTER between a machine translated document and its post-edited version, the way the post-editing was done makes it different from the standard HTER measure. The two-stage post-editing method emphasises document-level problems by ensuring that the translator concentrates on document-wide issues during the second stage of post-editing. Therefore, DISS-HTER is expected to be more informative in terms of document-level issues than a traditional HTER calculated against a corpus post-edited without following the two-stage post-editing method (such as the LIG corpus, presented in Section 5.4).

For the other labels, we combine the HTER score between PE1 and MT ($PE_1 \times MT$) with the HTER between PE2 and PE1 ($PE_2 \times PE_1$). This label aims to penalise documents with higher differences between PE1 and PE2, in other words, documents with more document-level problems. Labels penalising document-aware issues are expected to lead to better prediction models that reflect actual document-level quality.

In order to generate such a label, we use a linear combination of $PE_1 \times MT$ and $PE_2 \times PE_1$ (Equation 6.1), where w_2 and w_1 are empirically defined:

$$f = w_1 \cdot PE_1 \times MT + w_2 \cdot PE_2 \times PE_1, \tag{6.1}$$

For DISS-LC-P and DISS-LC-M, since the scale of $PE_1 \times MT$ is different from $PE_2 \times PE_1$, given that $PE_2 \times PE_1$ has much smaller values, we first normalised both distributions in order to make all values range between 0.0 and 1.0. The normalisation was done by applying Equation 6.2, where D is the set of labels, x is the data point to be normalised, x' is the normalised value, x_{min} is the minimum value of D and x_{max} is the maximum value of D.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}, \forall x \in D \tag{6.2}$$

[5]This work was done during my secondment at Pangeanic in Valencia, Spain http://www.pangeanic.com.

Following this normalisation, both HTER scores become comparable and they could be weighted accordingly . For **DISS-LC-P**, we varied w_1 from 0.0 to 1.0 ($w_2 = 1 - w_1$) and chose the value that maximises the Pearson r correlation score of a QE model. The QE models needed for learning the weigths were trained with the QUEST-17 features using the SVR algorithm in the `scikit-learn` toolkit, with the hyper-parameters (C, γ and ε) optimised via grid search. 10-fold cross-validation was applied in the data and the Pearson r scores are averages of all folds. In our experiment, $w_1 = 0.8$ and $w_2 = 0.2$ led to models with the best Pearson r score (after $1,000$ iterations).

For **DISS-LC-M**, the normalised data was also used and w_1 was also randomised between 0.0 and 1.0 but instead of maximising Pearson's r correlation scores we maximised the difference between $MAE_{predicted}$ and MAE_{mean}. QE models were built following the same configuration as for DISS-LC-P. Best results were achieved with $w_1 = 0.97$ and $w_2 = 0.03$ (after $1,000$ iterations).

DISS-LC-W2 was the quality label used in the WMT16 shared task on document-level QE. For this label, w_1 was fixed to 1.0, while w_2 was optimised to find how much relevance it should have in order to meet two criteria: (i) the final label (f) should lead to significant data variation in terms of the ratio between STDEV and the average (empirical maximum value of 0.5); (ii) the difference between the MAE_{mean} and $MAE_{predicted}$ should be maximised in each iteration. The data was not previously normalised as for DISS-LC-P and DISS-LC-M. QE models were built following the same configuration as for DISS-LC-P. The quality labels were defined by Equation 6.1 with $w_1 = 1.0$ and $w_2 = 13.0$ (the best w_2 was found after 20 iterations).

QE experiments and results

Features In the experiments reported in this section we used the following feature sets:

- QUEST-17;
- QUEST-ALL;
- QUEST-17+SHALLOW;
- QUEST-17+DEEP;
- QUEST-ALL+SHALLOW;
- QUEST-ALL+DEEP;

- QUEST-ALL+SHALLOW+DEEP;
- QUEST-17+WE;
- QUEST-17-CONSENSUS;
- ALL.

Method SVR and GP are used to generate the QE models. It is worth mentioning that SVR models were also used to optimise some of the new labels. Therefore, the labels can be biased towards this model. GP models aim to alleviate such impact, since it uses a completely different learning approach. For building the SVR models, we use the algorithm available in the `scikit-learn` toolkit with RBF kernel and hyper-parameters (C, γ and ε) optimised via grid search. For building the GP models, we use the GPy toolkit[6] with RatQuad kernels and the optimisation of hyperparameters done by maximising the model likelihood on the full training data

We followed the same split for training and test sets used in the document-level QE shared task on WMT16, where 146 documents were used for training the QE models and 62 for test. Table 6.7 shows the Pearson's r correlation scores for the models predicting either BLEU, TER or METEOR. Models predicting these metrics did not show significant Pearson's r correlation for the majority of the scenarios. The highest correlation scores for such models were a moderate correlation achieved when the WE feature set is used with SVR models, although for TER the model built with QUEST-ALL+SHALLOW+DEEP and SVR was not significant different from the model built with the WE feature set and SVR.

Table 6.8 show the Pearson's r correlation scores for the models predicting our new labels. All models predicting the new labels show significant Pearson's r correlation scores (with some exceptions when CONSENSUS features are used), with some scores falling into the strong correlation band. For DISS-HTER, the highest correlation scores were shown by the models built with QUEST-ALL+DEEP and GP, QUEST-ALL and GP, and QUEST-ALL+SHALLOW and SVR (no statistically significant difference was found between the models). Therefore, it seems that document-aware features are the most effective for this label. For DISS-LC-W2, the best models were built with QUEST-ALL and GP and ALL and GP. As for DISS-HTER, the document-aware features seem to be the most effective. For DISS-LC-P, the highest correlation was achieved by the models built with QUEST-ALL and GP, QUEST-ALL+DEEP and GP, and QUEST-17+WE and SVR. For this label document-aware features also seems to be creating the best models. Finally, for DISS-LC-M, all the best

[6]https://sheffieldml.github.io/GPy/

	BLEU		TER		METEOR	
	SVR	GP	SVR	GP	SVR	GP
QUEST-17	0.067*	0.057*	0.219*	0.125*	0.117*	0.088*
QUEST-ALL	0.044*	0.117*	0.113*	0.213*	0.039*	0.128*
QUEST-17+SHALLOW	0.119*	0.149*	0.038*	0.165*	0.059*	0.135*
QUEST-ALL+SHALLOW	0.013*	0.136*	0.031*	0.200*	0.007*	0.135*
QUEST-17+DEEP	0.193*	0.132*	0.252	0.220*	0.178*	0.176*
QUEST-ALL+DEEP	0.135*	0.179*	0.197*	0.284	0.131*	0.191*
QUEST-ALL+SHALLOW+DEEP	0.068*	0.199*	**0.289**	0.269	0.068*	0.197*
QUEST-17+WE	**0.325**	0.198*	**0.330**	0.283	**0.342**	0.259
QUEST-17+CONSENSUS	0.162*	0.040*	0.183*	0.098*	0.156*	0.073*
ALL	0.282	0.202*	0.297	0.197*	0.308	0.211*
RANDOM	0.046*		0.111*		0.006*	

Table 6.7 Results of different models predicting BLEU, TER and METEOR in terms of Pearson *r* correlation. * indicates results that did not show significant Pearson *r* correlation with p-value < 0.05. The best systems are highlighted in bold (William's significance test with p-value < 0.05).

models involve the DEEP feature set: QUEST-17+DEEP and SVR, QUEST-ALL+DEEP and SVR, and QUEST-ALL+DEEP and GP. Therefore, for this last label, predictions made with models built with discourse-aware features seem to be more reliable than predictions made with models without them. CONSENSUS features did not improve over baseline results (for the majority of the cases). Although the combination of different kernels for different feature types with GP did not improve over our single kernel SVR counterpart for QUEST-17+WE and QUEST-17+CONSENSUS, the multiple kernel models are consistently better for the ALL feature set than their SVR counterparts.

	DISS-HTER		**DISS-LC-W2**		**DISS-LC-P**		**DISS-LC-M**	
	SVR	GP	SVR	GP	SVR	GP	SVR	GP
QUEST-17	0.490	0.529	0.286	0.266	0.333	0.295	0.446	0.278
QUEST-ALL	0.531	**0.564**	0.309	**0.419**	0.481	**0.509**	0.475	0.485
QUEST-17+SHALLOW	0.503	0.479	0.297	0.264	0.481	0.337	0.442	0.337
QUEST-ALL+SHALLOW	**0.562**	0.523	0.267	0.344	0.419	0.457	0.443	0.496
QUEST-17+DEEP	0.531	0.555	0.218	0.245	0.405	0.448	**0.506**	0.443
QUEST-ALL+DEEP	0.556	**0.572**	0.324	0.373	0.488	**0.512**	**0.516**	**0.519**
QUEST-ALL+SHALLOW+DEEP	0.547	0.531	0.252	0.318	0.444	0.457	0.461	0.501
QUEST-17+WE	0.462	0.343	0.366	0.376	**0.510**	0.384	0.333	0.313
QUEST-17+CONSENSUS	0.475	0.428	0.225*	0.246*	0.422	0.229*	0.369	0.229*
ALL	0.349	0.398	0.360	**0.411**	0.423	0.449	0.343	0.399
RANDOM	0.046*		0.099*		0.090*		0.126*	

Table 6.8 Results of the models predicting our new labels for dissemination in terms of Pearson *r* correlation. * indicates results that did not show significant Pearson *r* correlation with p-value < 0.05. The best systems are highlighted in bold (William's significance test with p-value < 0.05).

Figure 6.2 shows results in terms of performance gains for models predicting BLEU, TER and METEOR. Models built with QUEST-17+WE with GP were the best for BLEU and METEOR, whilst the model with QUEST-ALL+SHALLOW+DEEP and GP was the best for TER. However, all gains are below 10% for all metrics.

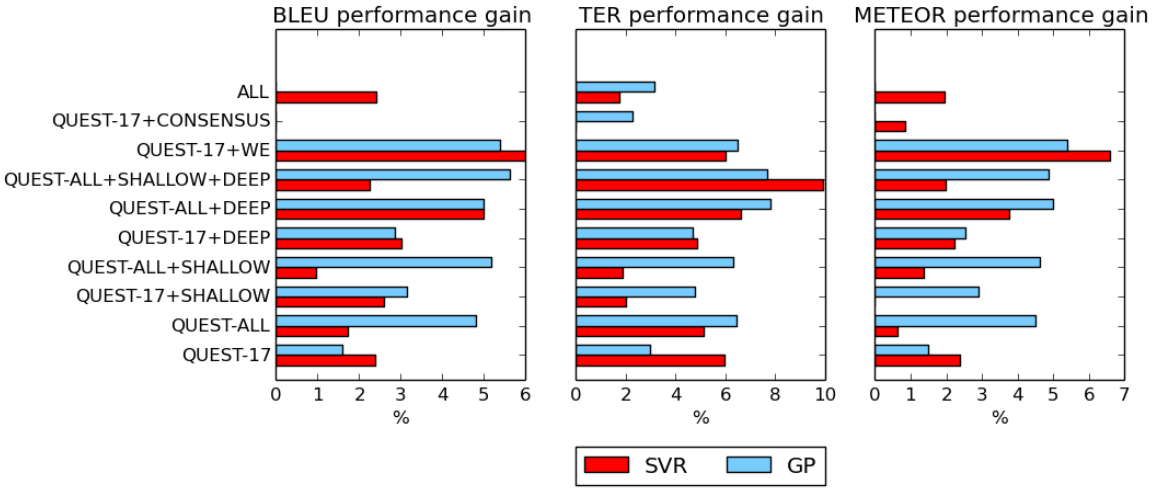

Fig. 6.2 Performance gains in terms of MAE of the QE models predicting BLEU, TER and METEOR.

Figure 6.3 shows the results of performance gains (in terms of MAE for the new dissemination labels). For DISS-HTER, the best models are built with SVR and the QUEST-ALL+SHALLOW+DEEP and QUEST-ALL feature sets. For DISS-LC-W2, the highest gains were achieved by the SVR model built with the QUEST+ALL features. The highest gains for DISS-LC-P is shown by both models using the QUEST-17+SHALLOW feature set. Finally, for DISS-LC-M, the best model was built with GP and the QUEST-ALL+DEEP feature set. All the highest performance gains were over 12% (the highest gain was achieved for HTER predictions - around 20% of performance gain).

The models built for the new labels show higher correlation scores and higher performance gains than the models predicting automatic metrics. In addition, the best models for our new labels are built with document and discourse-aware features, whilst the best models predicting BLEU-style metrics use WE features (with the exception of TER where the best model uses SHALLOW and DEEP features). Since our hypothesis is that our document and discourse-aware features should be a better representation from the documents than the other feature sets, we conclude that our new labels are more reliable in distinguishing documents according to document-level problems.

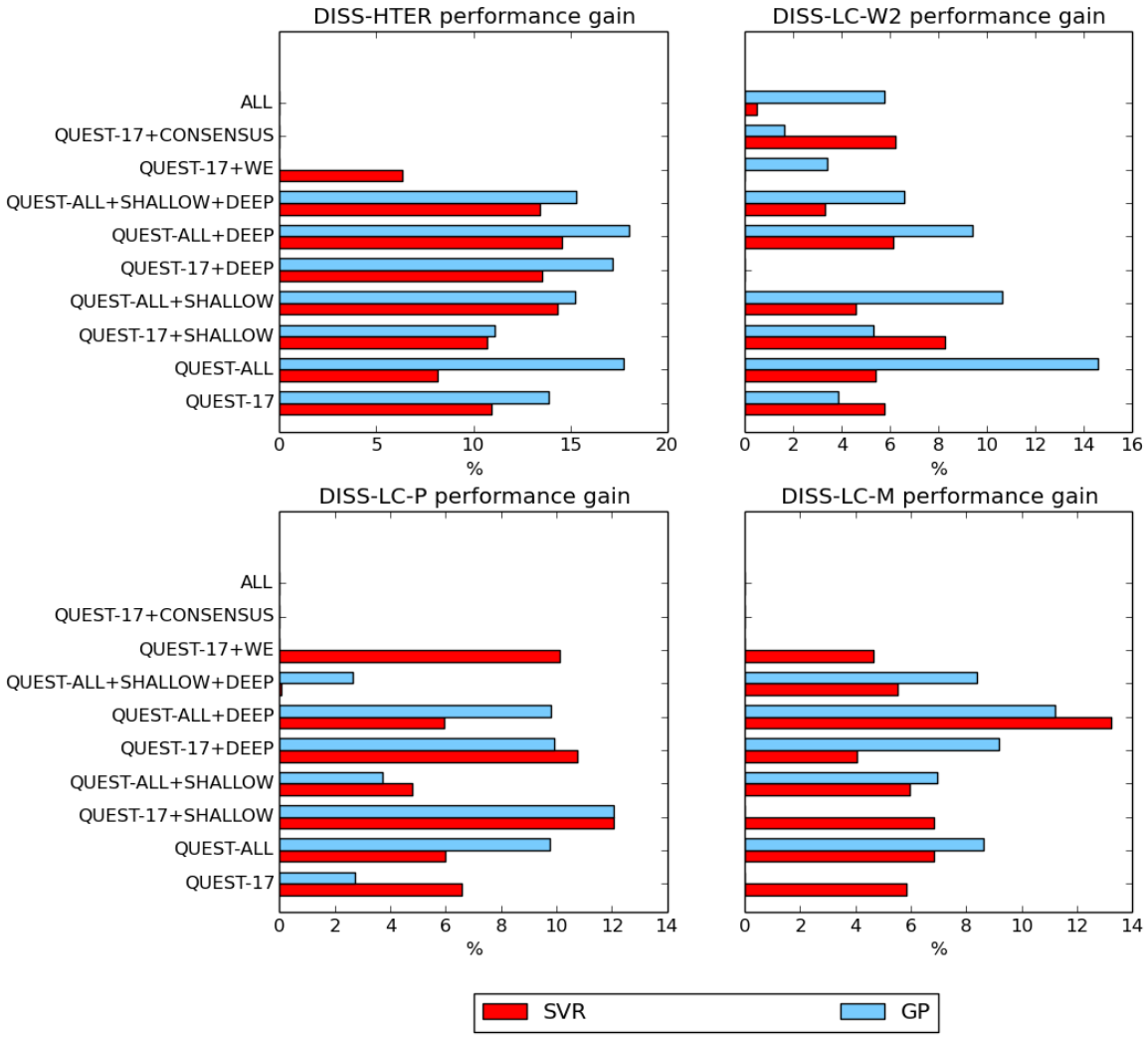

Fig. 6.3 Performance gains in terms of MAE of the QE models predicting the new dissemination labels.

Data Analysis

Table 6.9 shows the statistic dispersion and central tendency metrics for BLEU, TER, METEOR, DISS-HTER, DISS-LC-P and DISS-LC-M. Despite our efforts in maximising the variation of the data, the new labels do not show more variation than traditional automatic metrics. Moreover, the values for DISS-LC-W2 are not directly comparable to the other new labels and evaluation metrics, since the scale of the data is different. Looking at DISS-LC-W2 in isolation, the dispersion metrics do not show high variation on the data.

In Figures 6.4 and 6.5 we show the plots of the test set distributions along with the predictions made by the best systems for BLEU, TER and METEOR and our new quality

	MIN	MAX	MEAN	STDEV	Q_1	MEDIAN	Q_3	IQR
BLEU	0.060	0.681	0.233	**0.103**	0.159	0.219	0.295	**0.135**
TER	0.215	1.000	0.661	**0.137**	0.574	0.656	0.743	**0.170**
METEOR	0.234	0.819	0.468	**0.101**	0.395	0.462	0.530	**0.135**
DISS-HTER	0.171	0.645	0.381	**0.091**	0.317	0.370	0.434	**0.116**
DISS-LC-P	0.043	0.814	0.249	**0.107**	0.181	0.235	0.293	**0.113**
DISS-LC-M	0.006	0.975	0.252	**0.120**	0.178	0.231	0.301	**0.123**
DISS-LC-W2	0.180	2.706	0.895	**0.457**	0.602	0.765	1.081	**0.479**

Table 6.9 Statistic dispersion and central tendency metrics for all metrics and new labels derived from the two-stage post-editing method. Values in bold highlight the statistic dispersion metrics.

labels. For BLEU and METEOR, the data points are concentrated in a small range and the predictions fall in an even smaller range. TER is slightly more spreadout. However, for the new labels, the data points are also concentrated in a small range of the data. This can be explained by the fact that the new labels are derived from HTER, that also shows low data variation. Even DISS-LC-W2, that assumes a higher data spectrum has its data points concentrated in a small portion of such spectrum. Perhaps this kind of document-level evaluation (based on HTER) will always show low variation among documents. However, as we previously presented, predictions made by models using our discourse-aware features could better predict our new labels than automatic evaluation metrics.

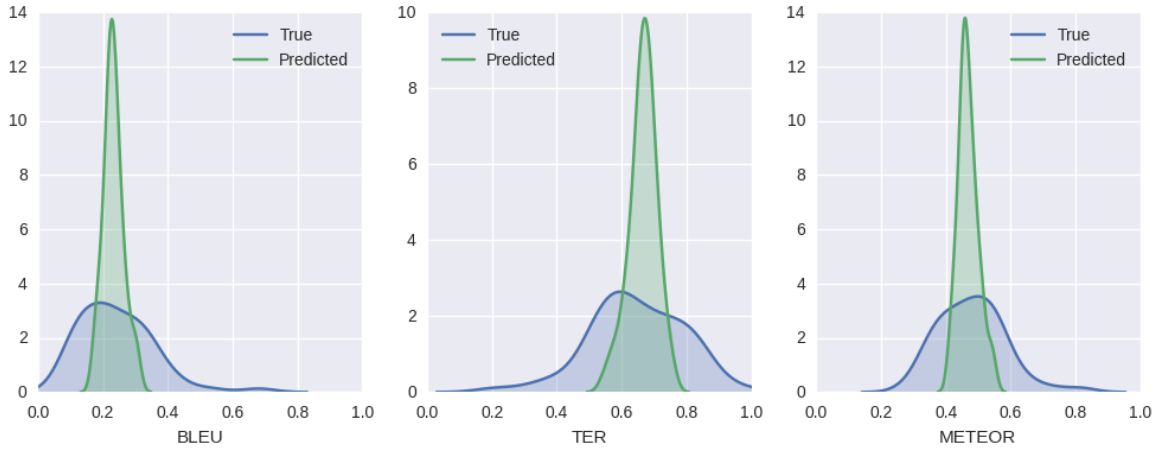

Fig. 6.4 Data distribution of true and predicted values of the best systems predicting BLEU, TER and METEOR.

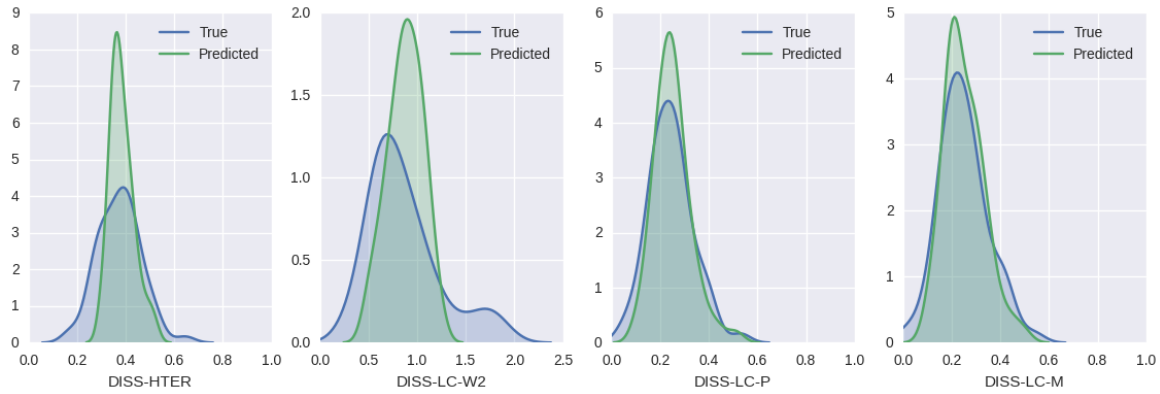

Fig. 6.5 Data distribution of true and predicted values of the best systems predicting DISS-HTER, DISS-LC-W2, DISS-LC-P and DISS-LC-M.

6.3 Assimilation Labels: Reading Comprehension Tests

One way to assess machine translated documents is by asking questions about these documents to humans. Reading comprehension questions aim at evaluating to what extent a human understands a given text. For that, the text does not need to be perfectly written, but its meaning must be understandable. Assuming that a group of fully literate speakers of a given language are able to answer questions regarding a given document, success or failure in the answers can be used to assess the quality of documents themselves. In MT evaluation, machine translated documents can be indirectly assessed by asking humans to answer reading comprehension questions about these documents (Jones et al., 2005b,a; Berka, Černý, and Bojar, 2011).

Reading comprehension questions can vary in form and complexity. Day and Park (2005) classify questions according to two dimensions: question forms and type of comprehension. Question forms define the kind of answer a question require (e.g. yes/no, multiple choice, *wh*-question). The type of comprehension is related to the effort required by the test taker in order to answer a question. Questions are then classified as "literal", "reorganization", "inference", "prediction", "evaluation" or "personal response".

Our experiments focus on two different corpora with reading comprehension questions: CREG and MCtest. CREG (Ott, Ziai, and Meurers, 2012) is a reading comprehension corpus for second language learners of German created with the aim of building and evaluating systems that automatically correct answers to questions. The texts were selected and the questions manually created by experts in the area of second language learner proficiency assessment. The corpus includes open questions that can be classified as "literal", "reorgani-

zation" or "inference" (Section 6.3.1). We use CREG-mt-eval, a version of CREG that we machine translated (Scarton and Specia, 2016). In this corpus, the questions are open and have different levels of complexity. Quality labels are derived by combining the marking of the questions answered by volunteers and information about question complexity. QE is then modelled as a regression task, as the quality scores per document vary following a continuous distribution. MCtest[7] (Richardson, Burges, and Renshaw, 2013) is a reading comprehension corpus built for benchmarking question answering systems. It contains multiple choice questions about English documents (Section 6.3.2). The quality labels are thus discrete and QE is approached as a classification task (we call our versions of MCtest, MCtest-mt-eval).

Statistics about the two corpora are summarised in Table 6.10.

	Documents	Words per document	Questions per document
CREG-mt-eval	108	372.38	8.79
MCtest-mt-eval	660	208	4

Table 6.10 Number of documents, average number of words and questions per document in CREG-mt-eval and MCtest-mt-eval corpora.

6.3.1 Experiments with CREG-mt-eval Corpus

The CREG-mt-eval corpus contains 108 source (German) documents from CREG with different domains, including literature, news, job adverts, and others (on average 372.38 words and 32.52 sentences per document). These original documents were translated into English using four MT systems: an in-house MOSES system (Koehn et al., 2007), Google Translate[8], Bing[9] and SYSTRAN.[10] Additionally, the corpus contains a version of each document with one sentence translated by each MT system (called 'mixed'). Finally, we included professional human translations of a subset of 36 documents as a control group to check whether or not the reading comprehension questions are adequate for the task.

An example of a document and its questions is presented in Table 6.11. A machine translation (Google) and a human translation are also presented in this table. It is possible to observe that, with only the MT output and/or only English knowledge, is very difficult (or impossible) to answer questions 2, 3 and 4.

The reading comprehension questions from CREG were translated by a professional translator. Questionnaires were generated with six translated documents (and their questions)

[7] http://research.microsoft.com/en-us/um/redmond/projects/mctest/
[8] http://translate.google.co.uk/
[9] https://www.bing.com/translator/
[10] http://www.systransoft.com/

Original: Objektbeschreibung einer 15-jährigen Wohneinheit Am Ende der Straße umgeben von Einfamilienhäusern erreichen Sie Ihr neues Heim. Von der Eingangstür treten Sie in den oberen Bereich dieser wunderbaren Wohneinheit, die die Eigentümer sehr sorgfältig und mit Liebe zum Detail removiert haben. Im Erdgeschoss befinden sich ein Bad mit Tageslicht, Gäste WC, die Küche und ein äußerst geräumiges Wohn/Esszimmer mit faszinierendem Blick auf den gepflegten Garten. Die Treppe hinunter sind ein weiteres Bad mit Dusche - bisher noch nicht benutzt - sowie zwei gleich große Räume, beide mit Blick auf den herrlichen Garten und das angrenzende Waldgebiet. Die Zimmer in diesem Bereich sind in hochwertigem Laminat ausgelegt. Wenn Sie verkehrsgünstig wohnen möchten und gleichzeitig eine familiäre Umgebung schätzen, ist diese Wohnung für Sie richtig. **Questions:** 1- Für wen ist diese Wohnung ideal? 2- Ist die Wohnung in einem Neubau oder einem Altbau? 3- Nennen Sie zwei Zimmer im Erdgeschoss. 4- Wo ist die Wohnung? 5- Wie viele Zimmer gibt es im Keller?
MT (Google): Description a 15-year residential unit At the end of the street surrounded by family houses you reach your new home. From the front door you enter into the upper region of this wonderful residential unit who remo four very carefully and with attention to detail the owners. Downstairs there is a bathroom with daylight, guest toilet, kitchen and an extremely spacious living / dining room with a fascinating view are the landscaped garden. The stairs are a further bathroom with shower - not yet used - and two equally sized rooms, both overlooking the beautiful garden and the adjacent forest. The rooms in this area are designed in high-quality laminate. If you want to stay conveniently and simultaneously appreciate a family environment, this apartment is right for you. **Questions:** 1- For whom is this apartment ideal? 2- Is the apartment in a new building or an old building? 3- Name two rooms on the ground floor. 4- Where is the apartment? 5- How many rooms are in the basement?
Human Translation: Property description for a 15-year-old residential unit Your new home is at the end of the street surrounded by single-family homes. When you enter the front door, you find yourself on the upper floor of this wonderful property which the owners have carefully renovated and decorated with much attention to detail. The ground floor has a bathroom with natural light, a guest toilet, the kitchen and a spacious living/dining room with a fascinating view of the beautiful garden. Downstairs you will find an additional bathroom with shower (that has not yet been used) and two equally large bedrooms overlooking the wonderful garden. The downstairs rooms have high-quality laminate flooring. If you want to enjoy the benefits of a convenient location with a suburbian flair, this property is perfect for you.

Table 6.11 Example of a document in the CREG corpus and its machine translation

each. Fluent speakers of English were then asked to answer sets with at least three questions about each document. Two versions of the questionnaires were build. The first had the six translated documents ordered as follows: MOSES, Google, Bing, SYSTRAN, 'mixed' and human. The second had the six translated documents ordered as: 'mixed', SYSTRAN, human, MOSES, Bing and Google. These two versions aim to maximise the number of different translations in the dataset. Questions about 216 documents (including 36 translated by a professional translator) were answered by test takers. It is worth noting that this set includes repetition of source documents, annotated by different users, but no repetition of target documents (i.e. they were translated by different systems). The questionnaires were answered using an online form by staff members and students from the University of Sheffield, UK.

The guidelines were similar to those used in reading comprehension tests: we asked the test takers to answer the questions using only the document provided. The original document (in German) was not given, therefore, test takers were not required to know German, but rather to speak fluent English. They were paid per questionnaire (set) and they were not able to evaluate the same set twice to prevent them from seeing the same document translated by a different system. Five sets were selected to be annotated five times, each time by a different test takers, so that agreement between them could be calculated.

Question Classification

As previously mentioned, the reading comprehension questions are open questions, and thus any answer could be provided by the test takers. Another important detail is that these questions have different levels of complexity, meaning that some questions require more effort to be answered. Since our aim is to generate quality labels from the answers, information about the questions' complexity level is important. We therefore manually classified the questions using the classes introduced by Meurers, Ott, and Kopp (2011), focusing on comprehension types (Day and Park, 2005).

Comprehension types: in order to identify the type of comprehension that a question encodes, one needs to read the text and identify the answer. The types of comprehension of the questions in CREG-mt-eval are:

- **Literal questions:** can be answered directly from the text. They refer to explicit knowledge, such as facts, dates, location, names.

- **Reorganisation questions:** are also based on literal text understanding, but the test taker needs to combine information from different parts of the text to answer these questions.

- **Inference questions:** cannot be answered only with explicit information from the text and involve combining literal information with world knowledge.

Question Marking

The most important aspect in the process of generating our document-level quality label is the correctness of the questions. In order to mark the answers to the questions, we follow the work of Ott, Ziai, and Meurers (2012), where the answers' classes, based on the gold-standard (target) answer(s), are the following. For each of these classes, we assigned numeric marks (in brackets):

- **Correct answer:** the answer is a paraphrase of the target or an acceptable answer for the question (score = 1.0).

- **Extra concept:** incorrect extra concepts are added to the answer (score = 0.75).

- **Missing concept:** important concepts of the answer are missing (score = 0.5).

- **Blend:** mix of extra concepts and missing concepts (score = 0.25).

- **Non-answers:** the answer is completely incorrect (not related to the target answer) (score = 0.0).

Table 6.12 shows the relative frequency of each marking category.

Grade	Mark	Frequency (%)
Correct	1.00	64.48
Extra concept	0.75	51.80
Missing concept	0.50	10.47
Blend	0.25	59.20
Non-answer	0.00	13.95

Table 6.12 Question grades, marks and frequency of the marks in CREG-mt-eval.

Test takers agreement

The agreement was calculated by using the Fleiss' Kappa metric. Alternatively, Spearman's ρ correlation coefficient was also calculated as the average between the ρ figure between each pair of test takers. Table 6.13 show results for Fleiss' Kappa and Spearman's ρ for the five sets.

	Scenario 1		Scenario 2	
	Fleiss' Kappa	Spearman's ρ	Fleiss' Kappa	Spearman's ρ
SET1	0.461	0.318	0.490	0.334
SET2	0.269	0.187	0.245	0.102
SET3	0.324	0.283	0.193	0.099
SET4	0.581	0.577	0.342	0.203
SET5	0.328	0.274	0.211	0.110

Table 6.13 Test takers agreement per set.

All sets except SET3 from *Scenario 2* show fair or moderate agreement according to Fleiss' Kappa. Spearman's ρ values are directly proportional to Fleiss. The best agreement is found in SET4 from *Scenario 1* (0.581 for Fleiss' Kappa and 0.577 for Spearman's ρ) and the worst in SET3 (0.269 and 0.187 for Fleiss' Kappa and Spearman's ρ, respectively).

We conducted further analyses on the data in an attempt to identify why some sets achieved worse results than others. Firstly, we hypothesised that sets with lower agreement figures could contain more difficult questions or, in other words, more questions classified as 'reorganisation' and 'inference'. However, this hypothesis proved false, since SET3 (*Scenario 2*) only has literal questions and SET4 (*Scenario 1*) has a uniform mix of all types of questions.

We also computed the correlation between the number of words in a set and its Fleiss' Kappa agreement. Table 6.14 shows the number of words and sentences per set. The correlation as calculated by Spearman's ρ was -0.60, indicating that when the number of words increases, the agreement decreases. However, it is worth noticing that SET3 from *Scenario 2*, which showed the worst agreement, is not the largest set in terms of words.

	Scenario 1 Number of words	Scenario 2 Number of words
SET1	2221	2230
SET2	3110	3152
SET3	2390	2391
SET4	2090	3937
SET5	2286	2343

Table 6.14 Number of words per set.

Table 6.15 shows Fleiss' Kappa values per document in all sets. Some documents show very low or no agreement, indicating that humans had problems answering questions for those documents. Although it would be expected that test takers should perform better when answering questions on human translated documents, such documents present low agreement in the majority of the sets (values in bold in Table 6.15).

	Scenario 1					Scenario 2				
	SET1	SET2	SET3	SET4	SET5	SET1	SET2	SET3	SET4	SET5
doc 1	0.242	1.000	0.492	0.447	0.541	0.321	−0.071	0.048	0.333	−0.034
doc 2	0.301	0.275	0.200	0.207	0.327	0.363	0.176	0.021	0.476	−0.046
doc 3	0.644	0.528	0.253	0.254	0.182	**0.492**	**0.242**	**0.317**	**0.764**	**0.135**
doc 4	0.373	0.107	0.113	0.185	0.231	0.452	0.083	0.294	0.156	0.083
doc 5	0.321	−0.010	0.527	0.663	0.063	0.803	0.312	0.439	0.015	0.182
doc 6	**0.500**	**0.000**	**0.040**	**0.000**	**0.044**	0.417	0.299	0.044	−0.046	0.638

Table 6.15 Test takers Fleiss' Kappa agreement per document. It is worth noticing that document 1 (doc 1) in SET1 is different from doc 1 in SET2, doc 1 in SET3 and so on. Values in bold highlight values for human translation.

Table 6.16 shows the average agreement per system, considering all machine translated documents (12 documents per system in total). This table also shows the quality of the MT systems in terms of BLEU and the average performance of the test takers in answering the questions for each system. MOSES is the system that obtained the highest agreement on average, followed by Bing. Although SYSTRAN shows the worse inter-annotator agreement and the worse BLEU score, there seems to be no correlation between system quality and agreement among annotators. For instance, the human translations only achieved 0.211 of agreement, whilst the best agreement score was 0.316. Therefore, since human translation showed lower agreement among annotators than MOSES, Bing and Google, we hypothesise that the inter-annotator agreement values is not only attributed to the system's quality. Instead, the agreements among annotators might be defined by other factors, such as motivation of annotators, genre of documents, order in which the document appear in the questionnaire, among others.

In addition, the performance of the test takers in answering the questions does not seem to correlate with the system performance or with the inter-annotator agreements. As expected, the human translations showed the averaged score of 0.801, followed by SYSTRAN, Mixed, Google, Bing and MOSES. Perhaps the high agreement for MOSES can be explained by the lowest score in performance: the test takers may have had the same problems while answering questions for MOSES. However, the low agreement showed by human translations can not be explained following the same argument, since the performance for human translations

is the best among the different "systems". More investigation needs to be done in order to define whether or not the system quality and test takers agreement have an impact on the test takers overall performance. More annotations for calculating agreements and more documents answered for each system could shed some light in this direction.

	Fleiss' Kappa average	BLEU	Test takers performance
MOSES	0.316	0.259	0.657
Bing	0.300	0.280	0.681
Google	0.221	0.313	0.750
Human	0.211	1.00	0.810
Mixed	0.180	0.245	0.752
SYSTRAN	0.167	0.120	0.761

Table 6.16 Average inter-annotator agreement, overall quality (in terms of BLEU) and overall test takers performance per system.

Deriving Quality Labels

One way of using answers to devise a quality labels is to simply average the marking scores for each document. However, our questions have different levels of complexity and such a simple combination would not reflect the difficulty of answering the different types of questions availabel.[11] Questions were then manually categorised following the types of questions used in (Meurers, Ott, and Kopp, 2011). Table 6.17 shows the relative frequency of each type of question.

Type of Question	Frequency (%)
Literal	78.65
Reorganization	12.05
Inference	9.30

Table 6.17 Types of question and their frequency in CREG-mt-eval.

The two new labels (RC-CREG-P and RC-CREG-M) derived from the marking of CREG-mt-eval are generated, for each document, using Equation 6.3.

$$f = \alpha \cdot \frac{1}{Nl} \sum_{k=1}^{Nl} lq_k + \beta \cdot \frac{1}{Nr} \sum_{k=1}^{Nr} rq_k + \gamma \cdot \frac{1}{Ni} \sum_{k=1}^{Ni} iq_k, \tag{6.3}$$

[11] We experimented with the simple average of all questions and it did not lead good results - the QE models used to predict this label did not show statistically significant results.

where Nl, Nr and Ni are the number of "literal", "reorganization" and "inference" questions, respectively, lq_k, rq_k and iq_k are real values between 0 and 1, according to the mark of question k, and α, β and γ are weights for the different types of questions.

The weights α, β and γ were optimised following two different approaches. For **RC-CREG-P**, we use random search (Bergstra and Bengio, 2012), aiming at maximising the Pearson r correlation between the QE model and the final true labels. At each iteration, α was chosen randomly, from the range $[0.0, 1.0)$. Another random value ϕ was chosen randomly (also from the range $[0.0, 1.0)$) in order to define β as $(1-\alpha)\cdot\phi$ and γ as $(1-\alpha)\cdot(1-\phi)$. A QE model was trained at each iteration and the Pearson r correlation values were computed. The QE models were trained with QUEST++ baseline features at document level (Specia, Paetzold, and Scarton, 2015). The QE models were trained with the SVR algorithm in the `scikit-learn` toolkit, with the hyper-parameters (C, γ and ε) optimised via grid search. 10-fold cross-validation was applied and the Pearson r scores were the average of all folds. After $1,000$ iterations, the weights found were $\alpha = 0.614$, $\beta = 0.370$ and $\gamma = 0.016$.

For **RC-CREG-M**, we use random search and aim to maximise the difference between MAE_{naive} and $MAE_{predicted}$. Similarly to the first approach, at each iteration, α was chosen randomly, from the range $[0.0, 1.0)$. Another random value ϕ was chosen randomly (also from the range $[0.0, 1.0)$) in order to define β as $(1-\alpha)\cdot\phi$ and γ as $(1-\alpha)\cdot(1-\phi)$. A QE model, trained with the same configuration for RC-CREG-P, was trained at each iteration and the difference between MAE_{naive} and $MAE_{predicted}$ was computed. After $1,000$ iterations, the weights found were $\alpha = 0.105$, $\beta = 0.619$ and $\gamma = 0.276$.

In RC-CREG-P, the highest weight was given for literal questions (α parameter), whilst in RC-CREG-M, the highest weight was given to reorganization questions (β parameter). Therefore, for RC-CREG-P, simple questions are more problematic: if the test taker could not answer these questions properly, the document will be more heavily penalised. On the other hand, for RC-CREG-M, questions with an intermediate difficulty are more problematic. Moreover, for this label, inference questions are also weighted higher than literal questions (the parameter γ is higher than α). This means that in RC-CREG-M, difficult questions are more important in generating the label than literal questions.

Table 6.18 shows the statistic dispersion and central tendency metrics for RC-CREG-P and RC-CREG-M. Values for IQR are slightly higher than the values shown for BLEU, TER and METEOR in Section 5.5 and the values for the dissemination labels (Table 6.9). However, this is a completely different task and data and, therefore, it is not possible to compare these results in a fair manner.

	MIN	MAX	MEAN	STDEV	Q_1	MEDIAN	Q_3	IQR
RC-CREG-P	0.167	0.833	0.571	**0.155**	0.459	0.583	0.685	**0.226**
RC-CREG-M	0.167	0.833	0.542	**0.165**	0.442	0.541	0.667	**0.225**

Table 6.18 Statistic dispersion and central tendency metrics for RC-CREG-P and RC-CREG-M.

QE Experiments and Results

Features In the experiments reported in this section we used the following feature sets: QUEST-17, QUEST-ALL, QUEST-17+SHALLOW, QUEST-ALL+SHALLOW, QUEST-17+WE and ALL. CONSENSUS features were not used in this experiment since the human translations were also mixed into the data set and we only have 36 human translated documents. DEEP features were also not available since the source language is German.

Method SVR and GP are used to generate the QE models. As for the dissemination labels, SVR models were also used to optimise some of the new labels. We expect to alleviate this bias by showing experiments with GP models as well as SVR models. Therefore, we can compare the results and understand the bias (if it exists).

We compare the models trained to predict RC-CREG-P and RC-CREG-M against those trained on automatic MT evaluation metrics as labels. Ideally, in order to show that our new label is more reliable than automatic metrics, human translated data would need to be used as reference for the automatic metrics. However, as mentioned before, only 36 documents from the CREG-mt-eval corpus are human translated. In order to build models for BLEU, TER and METEOR for comparison purposes, we sample data from a different corpus with the same language pair.

This corpus was extracted from the WMT 2008-2013 translation shared task corpora for DE-EN, and totals 474 documents (the same data used in Section 5.3). Although the datasets are different, our hypothesis is that they are comparable, given that the CREG-mt-eval corpus also contains news data.

We perform 10-fold cross-validation with both CREG-mt-eval and WMT corpora. Cross-validation was preferred over dividing the corpus into training and test sets, because this way we can study both corpora without the need of extra data manipulation. Moreover, CREG-mt-eval has a small number of documents, which could lead to overfiting. Results in terms of Pearson r correlation are shown in Table 6.19 (correlation scores are averaged over the 10-fold cross-validation iterations). Both RC-CREG labels achieved the highest correlation scores with models built using QUEST-17 and GP when compared to BLEU-style

metrics. However, the use of SHALLOW features did not appear to help in predicting the new labels. QUEST-17+WE features, on the other hand, obtained higher correlation scores, being the best for RC-CREG-P. Although models built with the ALL feature set predicting BLEU, TER, METEOR showed higher correlation scores than the best models for our new labels, such results cannot be interpreted as a drawback of our labels. The new labels proposed focus on a different task and use different data than the one used for predicting the automatic evaluation metrics.

	BLEU		TER		METEOR		**RC-CREG-P**		**RC-CREG-M**	
	SVR	GP	SVR	GP	SVR	GP	SVR	GP	SVR	GP
QUEST-17	0.254	0.266	0.253	0.286	0.286	0.237	0.209	0.342	0.267	0.343
QUEST-ALL	0.360	0.367	0.422	0.408	0.340	0.358	0.160	0.350	0.202	**0.385**
QUEST-17+SHALLOW	0.283	0.333	0.281	0.346	0.341	0.321	0.279	0.307	0.118	0.247
QUEST-ALL+SHALLOW	0.385	0.387	0.437	0.435	0.433	0.381	0.239	0.318	0.243	0.344
QUEST-17+WE	0.447	0.449	0.561	0.566	0.295	0.376	**0.476**	0.455	0.304	0.378
ALL	0.490	**0.493**	0.613	**0.639**	0.507	**0.507**	0.288	0.352	0.301	0.380

Table 6.19 Results in terms of Pearson *r* correlation of the models predicting the new RC-CREG labels and BLEU, TER and METEOR in the reference corpus. * indicates results that did not show significant Pearson *r* correlation with p-value < 0.05. The best systems are highlighted in bold (William's significance test with p-value < 0.05).

Figure 6.6 shows the performance gain of predicting BLEU, TER, METEOR, RC-CREG-P and RC-CREG-M. Although automatic evaluation metrics produced higher performance gains, the new labels also consistently yielded performance gains when QUEST-17, QUEST-ALL, WE and ALL features are used. For SHALLOW features, only small gains were achieved.

Finally, it is worth remembering that the WMT and CREG-mt-eval datasets are different and the results were just put together in order for us to have a reference for automatic evaluation metrics. Therefore, the results over WMT data should not be taken as an irrefutable evidence that such labels are better than ours. In fact, since our data suffers from high variance, we could argue that our labels are more reliable in distinguishing between documents. In summary, the use of marking from open questions as quality labels for QE of MT still needs further investigation. In terms of quality assessment, reading comprehension tests are reliable because they reflect the understanding and usefulness of a document for the end-users. However, the best way of devising labels for QE is still an open question.

Another issue with this corpus in particular is that the questions were developed for general purpose evaluation of the original documents. In this case, one side effect is that the parts of the documents needed to answer such questions may not be affected by the machine translation, which could mean that a very bad machine translation could still be

scored high. One solution would be to only use reading comprehension questions that always go beyond the sentence-level understanding (such as "reorganization" and "inference" questions). Unfortunately, this was not possible with the CREG corpus, for which some documents only have "literal" questions. Therefore, experiments with different data may show different findings, where labels devised from reading comprehension questions are actually significantly better than automatic evaluation metrics.

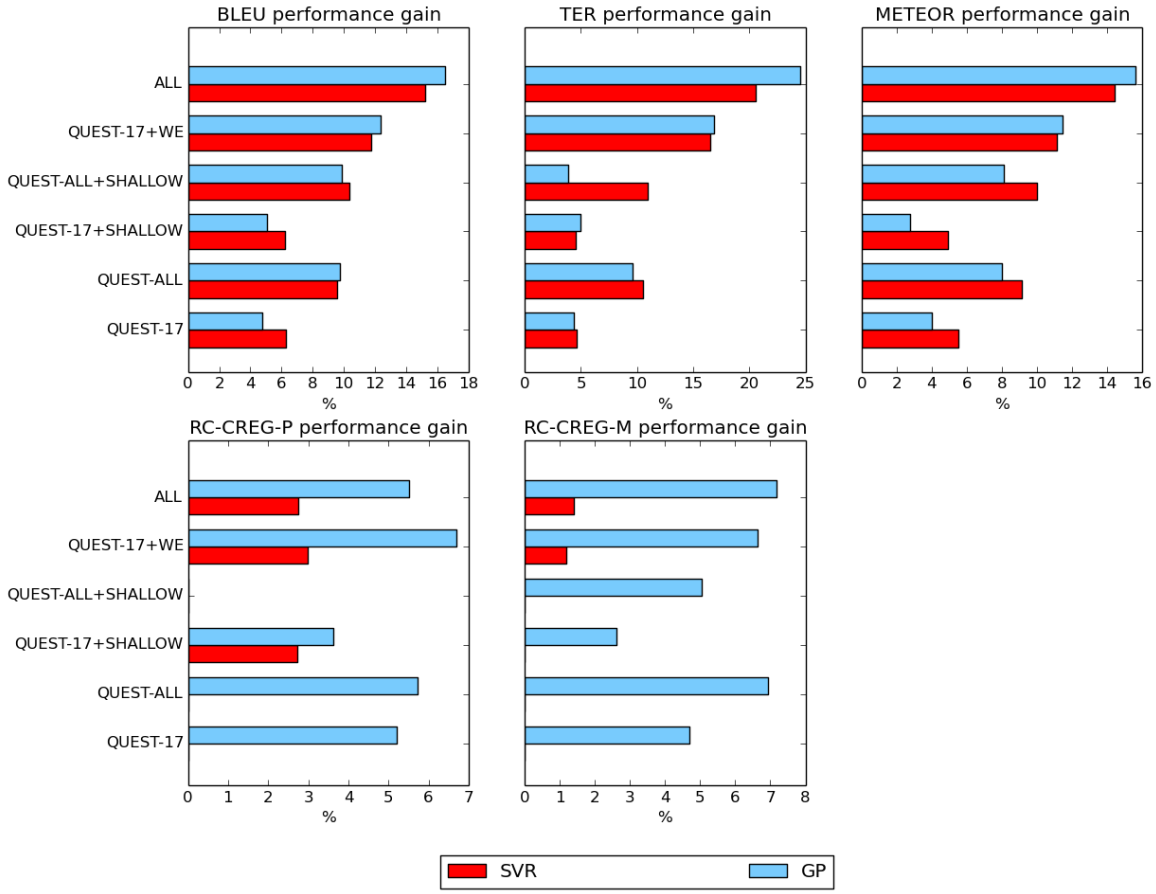

Fig. 6.6 Performance gains in terms of MAE of the models predicting the new RC-CREG labels and the automatic metrics in the reference corpus.

6.3.2 Experiments with the MCtest-mt-eval Corpus

The original MCtest corpus contains 660 documents in English with four multiple choice reading comprehension questions each. Since our test takers are native speakers of English, in order to use this corpus for MT evaluation, we first machine translated the English documents into German using a MOSES standard system (build with WMT15 data (Bojar

et al., 2015)), and then machine translated the German documents back into English. For the back translation we used a different MT system (BING translator) in order to maximise the chances of introducing translation errors of different types.

Questionnaires were built and test takers answered questions about one machine translated document and one (different) original document. Original documents were given to test takers as a control group. Since all questions are of the multiple choice type, the marking was done automatically and each answer was either marked as correct (1 mark) or incorrect (0 marks) (the processed MCtest corpus is called MCtest-mt-eval).

Deriving Quality Labels

The marks for a document vary between 0 (no questions answered correctly) and 4 (4 questions answered correctly). Given that all questions are considered equally complex in this corpus, no weights on questions were necessary. Figure 6.7(a) shows the distribution of correct answers in original documents versus machine translated documents. As expected, for original documents there is a higher frequency of documents with all questions answered correctly (4 marks) than for machine translated documents: 84% of original documents have all questions correctly answered, while only 52% of machine translated documents have all questions answered correctly.

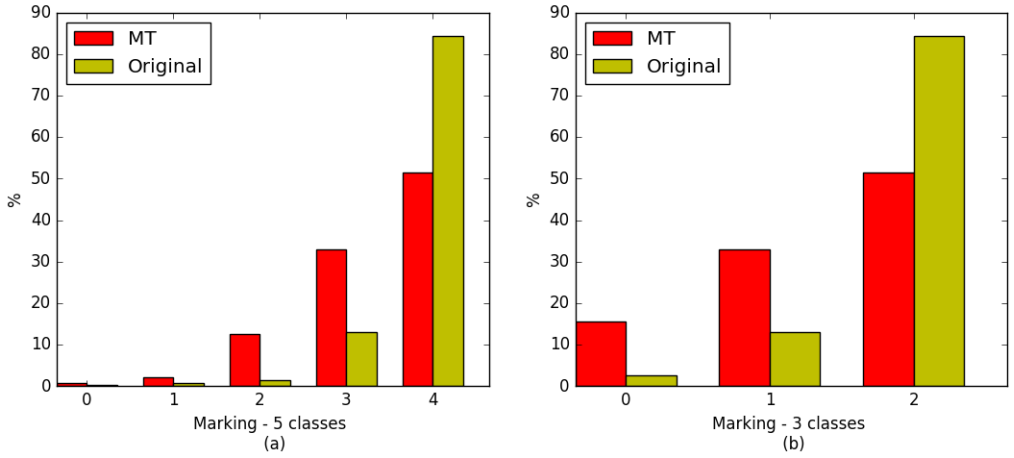

Fig. 6.7 Distribution of correct answers in original and machine translated documents in the MCtest corpus.

These marks are used directly as quality labels (this label is referred to hereafter as **RC-MCtest**). However, since there are only a few documents scoring 0 or 1, we combined these two classes with class 2. Therefore, we propose an approach with three options of

marking: 0, 1 and 2. The distribution of original and machine translated documents using these three classes is shown in Figure 6.7(b).

QE Experiments and Results

Features In the experiments reported in this section we used the following feature sets: QUEST-17, QUEST-ALL, QUEST-17+SHALLOW, QUEST-ALL+SHALLOW, QUEST-17+DEEP, QUEST-ALL+DEEP, QUEST-ALL+SHALLOW+DEEP, QUEST-17+WE and ALL. CONSENSUS features were not used in this experiment since we did not have many of the machine translations necessary. DEEP features were calculated by using the original documents in English.

Method Since the quality labels follow a discrete distribution, we addressed the problem as a classification task, instead of using regression. Two classification models were used: Random Forest and Ordinal Logistic. We use the Random Forest algorithm from `scikit-learn` and an Ordinal Logistic model from the mord toolkit[12] (Pedregosa-Izquierdo, 2015). Random Forests treat the problem as a multiclass classification task where the classes do not follow any order, while the Ordinal Logistic model is able to take the order of the classes into account. The Ordinal Logistic implementation follows the model proposed in (Rennie and Srebro, 2005). Since the labels have an order, ordinal classification is expected to be the most suitable approach.[13]

Data Here we used the official training, development and test set provided in the original corpus, with the development and training sets concatenated. The final training set has 450 documents, while the test set has 210 documents.

Baseline As a baseline, we use the majority class (MC) classifier (all test instances are classified as having the majority class in the training set).

Evaluation Precision, recall and F-measure are used to evaluate the performance of the QE models.

Table 6.20 shows the results, in terms of precision (P), recall (R) and F-measure (F), for the MCtest classification task using the three classes structure. Best results in terms of

[12] http://pythonhosted.org/mord/

[13] Although we also experimented with SVM classifiers (using the implementation from `scikit-learn`), they were outperformed by Random Forests and, therefore, we did not report their results.

F-measure are achieved when using QUEST-ALL+DEEP with the Ordinal Regression and QUEST-17+WE with Random Forest. The highest *F*-measure scores overall are achieved by Random Forests.

	Random Forest			Ordinal Logistic		
	P	R	*F*	P	R	*F*
QUEST-17	0.367	0.386	0.373	0.398	0.438	0.413
QUEST-ALL	0.393	0.438	0.409	0.425	0.424	0.402
QUEST-17+SHALLOW	0.366	0.395	0.369	0.372	0.414	0.390
QUEST-17+DEEP	0.352	0.410	0.378	0.439	0.443	0.423
QUEST-ALL+SHALLOW	0.445	0.462	0.450	0.426	0.429	0.406
QUEST-ALL+DEEP	0.434	0.471	0.443	0.431	0.429	0.414
QUEST-ALL+SHALLOW+DEEP	0.397	0.438	0.416	**0.442**	**0.452**	**0.434**
QUEST-17+WE	**0.462**	**0.505**	**0.474**	0.407	0.405	0.403
ALL	0.389	0.343	0.403	0.426	0.429	0.422
MC	0.284	0.533	0.371	-	-	-

Table 6.20 Results for the models performing a classification task on MCtest with three classes.

In Figure 6.8 we present the normalised confusion matrix[14] for the best models in the three classes scenario. The Random Forest model is better at predicting the dominant class (class 2), while the Ordinal Logistic model shows better results for the intermediate class (class 1). Both models fail to classify the class 0, which has less examples than the other two classes.

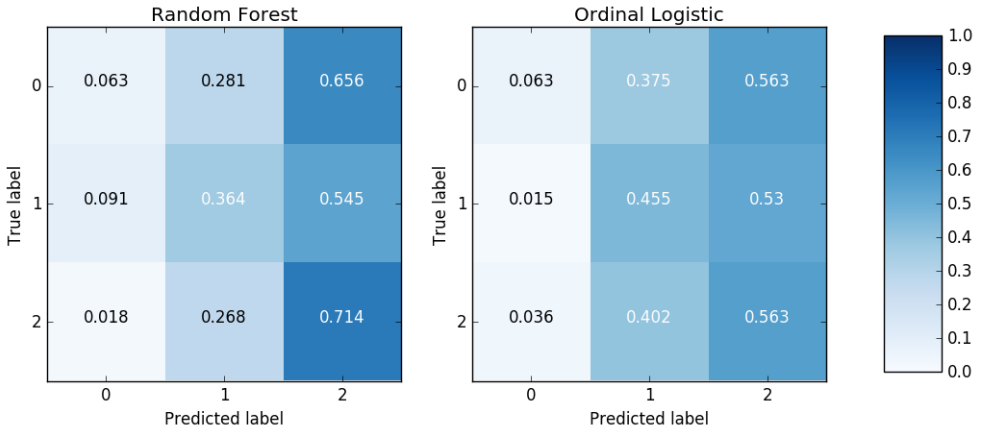

Fig. 6.8 Confusion matrix of the best models for the classification task on MCtest with three classes.

[14]The values are normalised by the number of elements in each class and, therefore, they can be seen as percentages.

Table 6.21 shows the results for the classification task using five classes. The best result overall is achieved by the model built with Random Forests and QUEST-ALL+DEEP feature set. The best result for Ordinal Logistic was achieved when using the ALL feature set.

	Five classes					
	Random Forest			Ordinal Logistic		
	P	R	*F*	P	R	*F*
QUEST-17	0.379	0.414	0.394	0.395	**0.443**	0.415
QUEST-ALL	0.360	0.400	0.371	0.368	0.410	0.385
QUEST-17+SHALLOW	0.433	0.452	0.442	0.403	**0.443**	0.417
QUEST-17+DEEP	0.403	0.462	0.422	0.398	0.433	0.410
QUEST-ALL+SHALLOW	0.437	0.476	0.447	0.375	0.424	0.397
QUEST-ALL+DEEP	**0.467**	**0.524**	**0.487**	0.392	0.438	0.412
QUEST-ALL+SHALLOW+DEEP	0.433	0.462	0.442	0.390	0.429	0.371
QUEST-17+WE	0.365	0.433	0.396	0.413	0.410	0.411
ALL	0.443	0.467	0.440	**0.433**	0.429	**0.428**
MC	0.284	0.533	0.371	-	-	-

Table 6.21 Results for the models performing a classification task on MCtest with five classes.

Figure 6.9 shows the normalised confusion matrix for the best models in the five classes scenario. Both models predict the dominant class (4) for the majority of the instances and fail to predict the correct labels for all instances in classes 0 and 1. The Ordinal Logistic model shows more variation in predicting the labels than the Random Forest model.

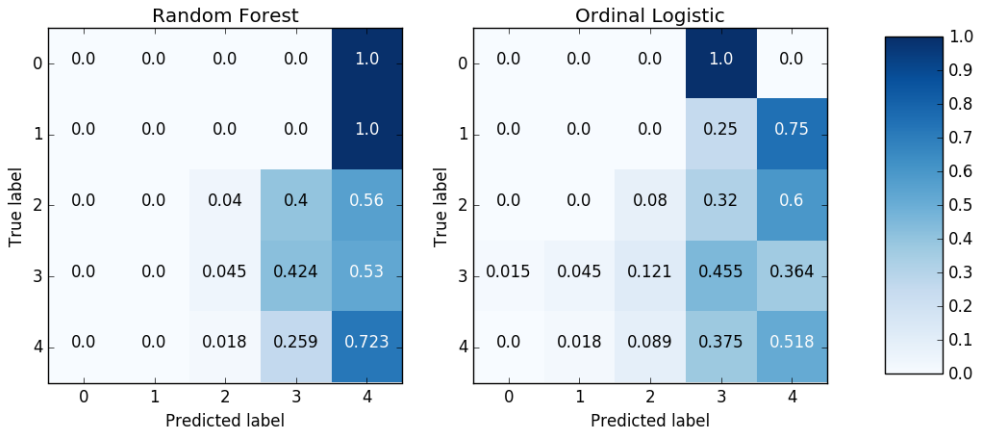

Fig. 6.9 Confusion matrix of the best models for the classification task on MCtest with five classes.

Using the original (untranslated) documents in the control group mentioned previously, the majority class would result in an *F*-measure of 0.764. This value is considerably higher than 0.487 or 0.474, which were found for the machine translated dataset (using five and

three classes respectively), indicating that the errors in our machine translated them made documents much more difficult to comprehend.

6.4 Discussion

In this chapter we presented our contribution in creating new quality labels for document-level QE. Our new labels are task-oriented and human-targeted and aim to assess documents more reliably than traditional MT evaluation metrics.

Section 6.1 featured our first experiments towards new quality labels for document-level QE. We showed that direct human assessments in terms of cohesion and coherence at document level did not lead to a reliable evaluation. For the majority of the sets, the agreement scores were very low, specially when assessing coherence. Humans were not able to isolate document-level problems and, therefore, the evaluation was unreliable. The two-stage post-editing method was also introduced in this section, with analyses on inter-annotator agreement and types of changes performed. This method showed promising results, with high agreement in the first stage of post-editing. The low agreement in the second stage is likely to be caused by the differences in the number of stylistic changes revealed by our manual analysis. Experiments in a controlled environment, performed by expert translators and with guidelines requesting minimal stylistic changes were expected to show more reliable results. We conclude that the new two-stage post-editing method is more viable and cost-effective than the direct human assessments.

In Section 6.2 we presented our large-scale experiments with the two-stage post-editing method aiming at dissemination purposes and compared the new labels with BLEU, TER and METEOR. Although our new quality labels did not achieve significantly higher data variance, models built with discourse-aware features for predicting such labels showed significant better performance than models built with baseline features. Whilst the automatic metrics did not produce reliable models (in terms of Pearson's r correlation) for the majority of the cases, the models built for the new labels consistently obtained significant Pearson's r correlation scores. Moreover, models built with discourse-aware features performed best for our new labels, but not for the automatic metrics. Therefore, using document-aware and discourse-aware features we were able to build reliable models for predicting our new labels. This is an evidence that our new labels are able to capture the document-wide issues that are addressed by our new features.

Section 6.3 presented our experiments for assimilation purposes using reading comprehension tests scores as a proxy for document quality. Two corpora were used and the

problem was approached as a classification (for multiple choice questions - MCtest) and a regression (for open questions - RC-CREG). For the RC-CREG labels, our regression models did not show higher correlation scores than models built for a reference corpus predicting BLEU, TER or METEOR. However a direct comparison between the two datasets is not possible, since the datasets are considerably different. Models predicting RC-CREG built with document-wide information showed moderate correlation with true labels. In addition, RC-CREG labels showed higher data variation than automatic evaluation metrics and, therefore, we can say that documents can be more easily distinguished with this label. This means that the new labels are also capturing some document-aware information. For MCtest labels, the best results were achieved by models built with document and discourse-aware features. Approaching the problem as a three-class task seemed to be the most reliable approach, given the labels distribution, although the best model (in terms of F-measure) was built for the five class scenario. Finally, the original and machine translated versions in MCtest were assessed according to their reading comprehension scores. We proved that machine translated documents are more difficult to comprehended than original documents. Thus, results showed that discourse-aware features perform well in this scenario. This means that MT is likely to be compromising the integrity of discourse structures necessary for document comprehension.

In summary, this chapter confirmed our hypothesis that task-based document-oriented evaluations of machine translated documents tend to be more reliable than automatic evaluation metrics. Models built with document and discourse-aware features were only competitive when predicting the new task-based labels proposed in this thesis. Moreover, task-based evaluation methods are more desired over general evaluation metrics because they provide a reliable assessment of the machine translated documents in terms of usefulness.

Chapter 7

Conclusions

This thesis presented our contributions for document-level QE in terms of feature engineering, quality labels and prediction models. Our literature review revealed a scarcity in previous work on document-level QE (Chapter 2). We also showed that recent work has focused on linguistic features for QE and promising results have been achieved for word and sentence-level tasks. Moreover, the literature review also evidenced the lack of studies in task-based assessments aimed at document-level QE. Task-based assessments are easy to interpret and desirable in order to evaluate MT output for a given purpose.

In our literature review we also discussed the use of document-aware information in MT, focusing on the use of discourse information (Chapter 3). Previous work on the use of discourse for improving MT, MT evaluation and QE motivated our discourse-aware features for document-level QE. Although such studies did not show substantial improvements over baselines, it is important to note the evaluation methodology used. The majority of work on MT is evaluated in terms of BLEU-style metrics, therefore, small but important changes in discourse may not be reflected in this type of evaluation. The study and application of new discourse-aware methods for MT are a challenge yet to be faced.

We introduced a new set of document and discourse-aware features designed for document-level QE (Chapter 4). Document-aware features encompass simple counts and aggregation of sentence-level features. Discourse-aware features use document-wide discourse information. We analysed the correlation of our new features and HTER, showing that discourse-aware features show correlation scores higher than the majority of document-aware features used as baselines. In our feature analysis we also show the potential of extracting discourse-aware features from the target documents, although the resources with discourse information are limited in this case. Finally, we also introduce the concept of consensus features based on pseudo-references. The consensus hypothesis is that if several MT systems output the

same translation, it is likely that such translation is correct. Although pseudo-references and, consequently, consensus achieved good results in previous work, such features should be used with caution. Pseudo-references are an unreliable and unavailable in realistic scenarios and, therefore, should be avoided.

Our new features were exhaustively explored in different datasets with different language pairs and machine translations (Chapter 5). We built document-level QE models with BLEU, TER and METEOR as quality labels for all datasets, with the LIG dataset also including HTER as quality labels. The performance of the QE models varied a lot across different datasets and language pairs. In fact, it was not possible to conclusively decide on which are the best ML techniques and feature sets. We also presented a discussion about the use of automatic evaluation metrics as quality labels for the document-level QE task. We show that there is no consensus on the data variation (for some datasets this variation was low, whilst for others it was considerably high). Therefore, it seems that automatic metrics are not reliable in assessing individual documents.

Finally, we proposed new methods for acquiring and devising new task-based document-aware quality labels for document-level QE (Chapter 6). Our first set of labels aimed to evaluate MT for dissemination purposes and used a two-stage post-editing method. Such method focused on isolating document-level problems from issues at other levels. New labels were devised through a linear combination of the different stages in post-editing. We showed that models built with discourse-aware features performed better when predicting the new labels when compared with predictions for BLEU, TER and METEOR. However, the data variation of the new labels was as low as the automatic evaluation metrics. Our hypothesis is that the new labels follow the data variation of HTER, given that they are derived from it, although they are able to assess documents in terms of document and discourse-aware issues.

The second set of labels we proposed were based on reading comprehension tests, focusing on MT for assimilation purposes. Reading comprehension tests about machine translated documents were answered by fluent speakers of the target language. The new labels were devised from the test markings. For RC-CREG, since the questions were open, the marking followed a continuous scale and, for MC-test, since the questions were multiple choice, the marking was done using a discrete scale. We approached RC-CREG as a regression task and MC-test as a classification task. Models predicting RC-CREG labels did not obtain high correlation scores and, moreover, such scores were not higher than the models predicting BLEU, TER or METEOR for a reference corpus. However, the differences between the two corpora may have benefited the reference corpus and, therefore, they are not directly comparable. Moreover, RC-CREG labels showed a high data variation,

which may be a sign that documents were distinctly assessed. For MC-test, models using discourse-aware features show good performance overall.

Our study on different ML techniques also did not show a clear tendency. Both SVR and GP achieved good results, which were highly dependent on the dataset used. Moreover, using different kernels for different sets of features did not achieve significant improvements over using a single kernel. Therefore, the contributions on this matter are unclear and more analysis needs to be done.

Despite our efforts, there are still open questions to be addressed for document-level QE. This is quite a new area of study and much more still needs to be done in terms of data analysis, features, labels and ML techniques. However, we expect that the findings provided in thesis will be used as a foundation for future work.

Document-level QE seems to be a viable task, although it is still unclear how to reliably assess documents. Given the low data variation presented by automatic evaluation metrics and our labels, perhaps the focus of document-level QE should be different. Instead of focusing on predicting the quality of an entire machine translated document directly, the task could focus on stages after some corrections or predictions were already performed. For example, document-level QE could be applied over a post-edited document, to guarantee that consistency and discourse are preserved. In this case, fine-grained problems would be already solved, making the domain of document-level QE more restricted.

Finally, a lesson learnt from this thesis is that quality labels should be deeply investigated and new alternatives should always be considered. The majority of the work in the QE area (for all levels) assumes that some quality labels (e.g. HTER) are ideal and base their results (in terms of feature engineering and/or ML model design) on the performance of models over these labels. However, more work investigating what such labels are capturing and whether or not they are adequate for a given task is still needed.

7.1 Future Work

Future work in document-level QE includes research on data acquisition, quality labels, feature engineering, ML, improving MT with QE and applying QE for task other than MT.

- **Data variety:** Data is known to be a problem for QE tasks. Since there is a need of labelled data, it is very difficult to acquire such data in large quantities and with the desired quality. Moreover, it is also known that QE models may differ for different domains. Therefore, there is an urgent need in acquiring data for document-level QE in order for us progress in this field.

- **Quality labels:** Although we introduce a first attempt in devising new task-based document-aware quality labels, this task is still a challenge. More studies in this topic are needed in order for us to clearly point out the best way of assessing documents for QE. For example, post-editing time could be a next quality label to be explored for the task. Another directions would be to explore the fact that sentences have different degrees of relevance inside a document. Therefore, the document quality could encompass the quality of sentences and their relevance. The hypothesis is that if sentences that are not very relevant for the understanding of the document show lower quality, this is less critical than a sentence with high relevance that has low quality.

- **Feature engineering:** With the acquisition of new data and new quality labels, more work on feature engineering will be needed. New data will bring new peculiarities to be explored that may be not addressed by the current features. Moreover, if large datasets are available with reliable quality labels, studies can be conducted in order to identify state-of-the-art features for the task. Features related to relevance of sentences can also be explored for scenarios where relevance is taken into account.

- **Deep learning approaches:** With the advent of new deep learning techniques that are able to extract information beyond words and sentences (e.g. Kenter, Borisov, and Rijke (2016)), QE at document level may evolve. Potentially the feature engineering step will no longer be needed and the focus will be on neural network architectures and their ability to learn different types of labels.

- **ML models:** Although the majority of work on QE uses previously developed and well-known ML techniques (e.g. SVM), there is a lack of studies on ML techniques specially designed and or modified for QE. There is no study of this kind for document-level QE. Another interesting topic is multi-level prediction, which could be approached in two ways: (i) using the predictions of fine-grained levels as features for document-level QE (e.g. Specia, Paetzold, and Scarton (2015)) or; (ii) using hierarchical models that consider the document as combination of fine-grained structures (e.g. Lin et al. (2015)).

- **Other NLP tasks:** The research presented in this thesis focused exclusively in document-level QE for MT. However, other NLP tasks could also benefit from our findings. For example, it is also a challenge to evaluate the results of Automatic Summarization and Text Simplification tasks, and QE approaches might be also interesting for them. Mainly for AS, where it important is to evaluate the entire document.

Appendix A

QUEST++ features

In this Appendix we present an exhaustive list of our new document and discourse-aware features implemented in the QUEST++ toolkit. We follow the same number order used in the tool, therefore, this appendix is also part of the QUEST++ documentation.

Document-aware features Features that were implemented based on sentence-level features:

1. **DocLevelFeature1001:** number of tokens in the source document;
2. **DocLevelFeature1002:** number of tokens in the target document;
3. **DocLevelFeature1003:** ratio between the number of tokens in the source document and in the number of tokens in target document;
4. **DocLevelFeature1004:** ratio between the number of tokens in the target document and in the number of tokens in source document;
5. **DocLevelFeature1005:** absolute difference between the number of tokens in the source document and the number of tokens in target document, normalised by the source length;
6. **DocLevelFeature1006:** average token length in the source document;
7. **DocLevelFeature1009:** LM log probability in the source document (values are averaged over sentence-level log probabilities);
8. **DocLevelFeature1010:** LM perplexity in the source document (values are averaged over sentence-level perplexity values);

9. **DocLevelFeature1011:** LM perplexity in the source document without end of sentence markers (values are averaged over sentence-level perplexity values);

10. **DocLevelFeature1012:** LM log probability in the target document (values are averaged over sentence-level log probabilities);

11. **DocLevelFeature1013:** LM perplexity in the target document (values are averaged over sentence-level perplexity values);

12. **DocLevelFeature1014:** LM perplexity in the target document without end of sentence markers (values are averaged over sentence-level perplexity values);

13. **DocLevelFeature1015:** Type/token ration for the target document;

14. **DocLevelFeature1016:** average number of translations per source word in the source document (threshold in Giza: prob $>$ 0.01);

15. **DocLevelFeature1018:** average number of translations per source word in the source document (threshold in Giza: prob $>$ 0.05);

16. **DocLevelFeature1020:** average number of translations per source word in the source document (threshold in Giza: prob $>$ 0.10);

17. **DocLevelFeature1022:** average number of translations per source word in the source document (threshold in Giza: prob $>$ 0.20);

18. **DocLevelFeature1024:** average number of translations per source word in the source document (threshold in Giza: prob $>$ 0.50);

19. **DocLevelFeature1026:** average number of translations per source word in the source document (threshold in Giza: prob $>$ 0.01) weighted by the frequency of each word in the source corpus;

20. **DocLevelFeature1028:** average number of translations per source word in the source document (threshold in Giza: prob $>$ 0.05) weighted by the frequency of each word in the source corpus;

21. **DocLevelFeature1030:** average number of translations per source word in the source document (threshold in Giza: prob $>$ 0.10) weighted by the frequency of each word in the source corpus;

22. **DocLevelFeature1032:** average number of translations per source word in the source document (threshold in Giza: prob > 0.20) weighted by the frequency of each word in the source corpus;

23. **DocLevelFeature1034:** average number of translations per source word in the source document (threshold in Giza: prob > 0.50) weighted by the frequency of each word in the source corpus;

24. **DocLevelFeature1036:** average number of translations per source word in the source document (threshold in Giza: prob > 0.01) weighted by the inverse of the frequency of each word in the source corpus;

25. **DocLevelFeature1038:** average number of translations per source word in the source document (threshold in Giza: prob > 0.05) weighted by the inverse of the frequency of each word in the source corpus;

26. **DocLevelFeature1040:** average number of translations per source word in the source document (threshold in Giza: prob > 0.10) weighted by the inverse of the frequency of each word in the source corpus;

27. **DocLevelFeature1042:** average number of translations per source word in the source document (threshold in Giza: prob > 0.20) weighted by the inverse of the frequency of each word in the source corpus;

28. **DocLevelFeature1044:** average number of translations per source word in the source document (threshold in Giza: prob > 0.50) weighted by the inverse of the frequency of each word in the source corpus;

29. **DocLevelFeature1046:** average unigram frequency in 1st quartile of frequency in the corpus of the source document;

30. **DocLevelFeature1047:** average unigram frequency in 2nd quartile of frequency in the corpus of the source document;

31. **DocLevelFeature1048:** average unigram frequency in 3rd quartile of frequency in the corpus of the source document;

32. **DocLevelFeature1049:** average unigram frequency in 4th quartile of frequency in the corpus of the source document;

33. **DocLevelFeature1050:** average bigram frequency in 1st quartile of frequency in the corpus of the source document;
34. **DocLevelFeature1051:** average bigram frequency in 2nd quartile of frequency in the corpus of the source document;
35. **DocLevelFeature1052:** average bigram frequency in 3rd quartile of frequency in the corpus of the source document;
36. **DocLevelFeature1053:** average bigram frequency in 4th quartile of frequency in the corpus of the source document;
37. **DocLevelFeature1054:** average trigram frequency in 1st quartile of frequency in the corpus of the source document;
38. **DocLevelFeature1055:** average trigram frequency in 2nd quartile of frequency in the corpus of the source document;
39. **DocLevelFeature1056:** average trigram frequency in 3rd quartile of frequency in the corpus of the source document;
40. **DocLevelFeature1057:** average trigram frequency in 4th quartile of frequency in the corpus of the source document;
41. **DocLevelFeature1058:** percentage of distinct unigrams seen in the corpus source (in all quartiles);
42. **DocLevelFeature1059:** percentage of distinct bigrams seen in the corpus source (in all quartiles);
43. **DocLevelFeature1060:** percentage of distinct trigrams seen in the corpus source (in all quartiles);
44. **DocLevelFeature1061:** average word frequency of the source document;
45. **DocLevelFeature1074:** percentage of punctuation marks in source document;
46. **DocLevelFeature1075:** percentage of punctuation marks in target document;
47. **DocLevelFeature1083:** percentage of content words in the target document;
48. **DocLevelFeature1084:** percentage of content words in the source document;

49. **DocLevelFeature1085:** ratio of the percentage of content words in the source document and the percentage of content words in the target documents;

50. **DocLevelFeature1086:** LM log probability of POS of the target document (values are averaged over sentence-level log probabilities values);

51. **DocLevelFeature1087:** LM perplexity of POS of the target document (values are averaged over sentence-level perplexity values);

52. **DocLevelFeature1088:** percentage of nouns in the source document;

53. **DocLevelFeature1089:** percentage of verbs in the source document;

54. **DocLevelFeature1090:** percentage of nouns in the target document;

55. **DocLevelFeature1091:** percentage of verbs in the target document;

56. **DocLevelFeature1092:** ratio of the percentage of nouns in the source document and the percentage of nouns in target document;

57. **DocLevelFeature1093:** ratio of the percentage of verbs in the source document and the percentage of nouns in target document;

58. **DocLevelFeature1300:** the Kullback-Leibler divergence between a source document and a target document topic distribution;

59. **DocLevelFeature1301:** the Jensen-Shannon divergence between a source document and a target document topic distribution;

60. **DocLevelFeature9300:** PCFG parse log likelihood of source documents (values are averaged over sentence-level values);

61. **DocLevelFeature9301:** average of the PCFG confidence of all possible parse trees in n-best list for the source document (values are averaged over sentence-level values);

62. **DocLevelFeature9302:** PCFG confidence of best parse tree for the source document (values are averaged over sentence-level values);

63. **DocLevelFeature9303:** number of possible PCFG parse trees for the source document (values are averaged over sentence-level values);

64. **DocLevelFeature9304:** PCFG parse log likelihood of target documents (values are averaged over sentence-level values);
65. **DocLevelFeature9305:** average of the PCFG confidence of all possible parse trees in n-best list for the target document (values are averaged over sentence-level values);
66. **DocLevelFeature9306:** PCFG confidence of best parse tree for the target document (values are averaged over sentence-level values);
67. **DocLevelFeature9307:** number of possible PCFG parse trees for the target document (values are averaged over sentence-level values);
68. **DocLevelFeature9800:** number of sentences in the source document;
69. **DocLevelFeature9801:** number of sentences in the target document;

Discourse-aware features Features that explore word repetition:

1. **DocLevelFeature9988:** content word repetition in the target document: the number of words that repeat are normalised by the total number of content words;
2. **DocLevelFeature9989:** content word lemma repetition in the target document: the number of lemmas that repeat are normalised by the total number of content words;
3. **DocLevelFeature9990:** content word repetition in the source document: the number of words that repeat are normalised by the total number of content words;
4. **DocLevelFeature9991:** content word lemma repetition in the source document: the number of lemmas that repeat are normalised by the total number of content words;
5. **DocLevelFeature9992:** ratio of content word repetition between the target and source documents;
6. **DocLevelFeature9993:** ratio of content word lemma repetition between the target and source documents;
7. **DocLevelFeature9994:** noun repetition in the target document: the number of words that repeat are normalised by the total number of content words;
8. **DocLevelFeature9995:** noun repetition in the source document: the number of lemmas that repeat are normalised by the total number of content words;

9. **DocLevelFeature9996:** ratio of noun repetition between the target and source documents;

Appendix B

Guidelines for quality annotation of MT outputs at paragraph level: discourse errors

B.1 Presentation

In this material we present the guidelines for quality annotation of MT outputs at paragraph level. These annotations will be done by humans, students from the "Translation Studies" courses at the Saarland University (Saarbrücken, Germany) and they are expect to reflect discourse problems among paragraphs. The students will receive certain paragraphs and will be asked to assess the quality of each individual paragraph.

B.2 Definitions

In this experiment, we define quality at paragraph level in terms of discourse quality. The phenomena under investigation are:

- Coherence: are the sentences of a given paragraph well connected? Does this contribute to keep the paragraph coherent? How much of the source paragraph coherence is found in the translation? The scores vary from 1 to 4.

    1. Completely coherent
    2. Mostly coherent
    3. Little coherent

4. Non-coherent

- Cohesion: are the cohesive devices correctly translated in a given paragraph? If not, do these errors change the information in the translation? (in other words, is it possible to interpret the paragraph correctly despite cohesion problems?) How appropriate are the cohesive devices in the machine translated paragraph? The scores vary from 1 to 4.

  1. Flawless (perfect, no errors)
  2. Good (minor errors that do not affect the understanding)
  3. Dis-fluent (major errors that make communication possible, but make the paragraph dis-fluent)
  4. Incomprehensible

The cohesive devices that should be looked for are:

- Reference: evaluate whether personal, demonstrative, possessive, relative, reflexive and indefinite pronouns are translated correctly and, if not, evaluate if this impacts the coherence of the paragraph.
- Connectives: evaluate whether connectives of the right category are used: Expansion, Contingency, Comparison, Temporal. The main idea is to evaluate whether the MT system choses the right (or the best) translation for a given connective.
- Ellipsis: where there are ellipsis, are they correctly translated?

B.3 Examples

B.3.1 Coherence

The coherence concept is more subjective than cohesion. A text is coherent if the sentences and paragraphs are connected in a logical way (Stede, 2011). In this activity, you should look for clues of disconnected ideas that lead to bad translations. To help you in this task, you will be provided with the source document. We ask you to first read the source document (in English) to identify the main idea and the connections between the sentences and paragraphs. After that, you are asked to read the machine translation (in German) and evaluate if the main ideas remain or if there are problems in logical structure of the paragraph. These problems can be signalised by wrong use of cohesive devices or mis-translation of a word or phrase,

changing the meaning of the paragraph or making it less logical. To explain the diferences between cohesion and coherence, let's consider an example of paragraph that is coherent and cohesive:[1]

*"My favourite colour is **blue**. I like it because **it** is calming and **it** relaxes me. I often go outside in the summer and lie on the grass and look into the **clear sky** when I am **stressed**. **For this reason**, I'd have to say my favourite colour is blue."*

This paragraph is coherent because the sentences make sense together (the person likes blue, because it is calm and relaxing). It is also cohesive, since the connections (highlighted) are made. It is also possible to have a cohesive paragraph with no coherence:

*"My favourite colour is **blue**. **Blue** sports cars go **very fast**. Driving **in this way** is dangerous and can cause many **car crashes**. I had a **car accident** once and **broke my leg**. I was very sad because I had to miss a holiday in Europe **because of the injury**."*

This paragraphs shows lots of cohesion devices and the sentences are well connected. However, the paragraph makes no sense: it is just a bunch of sentences describing different things connected together. Finally, one could also find a coherent paragraph with no cohesion:

"My favourite colour is blue. I'm calm and relaxed. In the summer I lie on the grass and look up."

As in the first example, this paragraph is showing a statement, a cause and a example, although it is more difficult to infer this logical meaning without the cohesive devices. However, one can still make logical connections mentally in order to understanding this paragraph. On the cohesion view, the paragraph is not cohesive since there are no connection among the sentences.

Some examples extracted from the corpora under investigation (translations from Spanish into English):

Example 1

[1]Examples extracted from: http://gordonscruton.blogspot.ca/2011/08/what-is-cohesion-coherence-cambridge.html

MT: In studies US detected a great index of infection between the groups sample **[]** it is hard interpret these data and make recommendations firm. The European study showed mortality certain difference among **patients had made detection and the not**.

SOURCE: En los estudios realizados en los Estados Unidos se detectó un gran índice de contagios entre los grupos de muestra, **por lo** que es difícil interpretar esos datos y hacer recomendaciones firmes. Por su parte, el estudio europeo mostró cierta diferencia de mortalidad entre **los pacientes** que se habían realizado la detección y **los** que no.

REFERENCE: In studies conducted in the United States, there was a lot of contamination between control groups, **so** it is difficult to interpret the data and make firm recommendations. Another study, this time a European one, concluded that there was a difference in mortality between patients **who** were screened and those **who** were not.

This example could be scored as “3 - Little coherent", since the absence of connectives or pronouns has several impacts in coherence, but one can still infer the logical relations.

Example 2

MT: Do **Hacerse** the test or not? Have asked board two specialists.

SOURCE: ¿**Hacerse** el test o no? Hemos pedido consejo a dos especialistas.

REFERENCE: Take the test or not? We asked two specialists for their opinion.

This example could be scored as “4 - Non-coherent", since a speaker of English will not understand the question.

B.3.2 Cohesion

Cohesion can be identified by superficial clues on the text. Different from coherence, it is less subjective in the sense that we should look for cohesion devices that are well marked along the document (Stede, 2011). Some examples are given as follows:

Examples on wrong usage of pronouns

MT: “Today, many men who have been detected a cancer are not treated, because the cancer is not aggressive and poses a risk to their lives. Instead, **they** suggest an active

surveillance of the disease and, if it progresses, **we** offer a treatment." == according to the source text, both should be “we" (possible score: 2 - Good)

SOURCE: “Hoy en día muchos hombres a los que se les ha detectado un cáncer no son tratados, puesto que dicho cáncer no es agresivo ni entraña un riesgo para su vida. En su lugar, **les** sugerimos una vigilancia activa de la enfermedad y, si ésta progresa, **les** ofrecemos un tratamiento").

MT: “My greatest wish is that **I cure** diarrhea, it is humiliating,’ he says. A few hours later, the team found to remedy this evil." == according to the source text the correct translation should be “to be cured of my diarrhoea" (possible score: 3 - Dis-fluent)

SOURCE: “Mi mayor deseo es que me curen la diarrea, es humillante’, confiesa. Unas horas más tarde, el equipo encontró remedio a ese mal."

MT: “The biggest fear of **Mrs A**, 89 years, is dying "conscious and drowned out." But the disease has made **him** to discover their children." == the correct pronoun is “her". (possible score: 2 - Good)

SOURCE: “El mayor temor de la Sra. A., de 89 años, es morir "consciente y ahogada". Pero la enfermedad **le ha** hecho descubrir a sus hijos. "Tengo unos buenos hijos", añade."

MT: “Now, the Brennan Center considers this a myth and claims that the electoral fraud is less common in the United States **that** the number of people who die as a result of the fall of lightning." == the right use is “than". (possible score: 2 - Good)

SOURCE: “Ahora bien, el Centro Brennan considera esto último un mito y afirma que el fraude electoral es menos frecuente en los Estados Unidos **que** el número de personas que mueren a causa de la caída de un rayo."

Examples on wrong usage of connectives

MT: “**Of fact**, lawyers Republicans have not found more 300 cases of electoral fraud United States ten years." == the right use should be “in fact" (possible score: 2 - Good)

SOURCE: “**De hecho**, los abogados republicanos no han encontrado más que 300 casos de fraude electoral en los Estados Unidos en diez años."

MT: “**Furthermore**, legislators Republicans supported 2011 certain laws abolishing entered constituents the same day of scrutiny eight States. **Furthermore**, cramping the right

of people and groups create a service of assistance voters to register. These restrictions have consequences." == according to the source text the first occurence of “furthermore" should be “on the other hand" (possible score: 3 - Dis-fluent)

SOURCE: “**Por otro lado**, los legisladores republicanos apoyaron en 2011 ciertas leyes que abolían la inscripción de electores el mismo día del escrutinio en ocho estados. **Además**, limitaron el derecho de las personas y los grupos a crear un servicio de ayuda a los electores que quisieran inscribirse. Estas restricciones tienen consecuencias."

B.4 Task description

You will receive a set of paragraphs (randomly selected from documents) and you are required to, for each paragraph, first read the source version (in English) and then read the machine translated paragraph (in German) twice: first looking for coherence problems and second looking for cohesion problems. For each paragraph, in each iteration, you are required to score the paragraph, appropriately for coherence and cohesion, according to the above definition of the 4-point scale.

The paragraphs are extracted from the WMT13[2] translation shared task test set. These documents are news texts in English that were machine translated into German.

You are required to fill a metadata form with the following format in which you should provide the evaluation scores for coherence and cohesion respectively:

<paragraph number_doc= number_total= number_sent=><doc_number><system>- ***coherence=***

<paragraph number_doc= number_total= number_sent=><doc_number><system>- ***coheresion=***

The parameters marked with “<" and “>" are headers to identify each single paragraph (you will observe that this header is the same that appear in the document with the paragraphs). The first parameter is related to the paragraph itself: “number_doc" is the order of the paragraph inside a given document; “number_total" is the oder of the paragraph considering the whole corpus; and “number_sent" is the number of sentences that this paragraph contains. The parameter “doc_number" refers to the number of the document where this paragraph appears and “system" is the machine translated system which produced this paragraph. The information follwing the header is the evaluation level: coherence or cohesion.

[2] http://www.statmt.org/wmt13/

As an example, consider that one student is evaluating the paragraphs:

<paragraph number_doc=2 number_total=92 number_sent=6><doc_6><rbmt-1_en-de_2013>-coherence=

<paragraph number_doc=2 number_total=92 number_sent=6><doc_6><rbmt-1_en-de_2013>-cohesion=

<paragraph number_doc=2 number_total=200 number_sent=4><doc_13><uedin_en-de_2013>-coherence=

<paragraph number_doc=2 number_total=200 number_sent=4><doc_13><uedin_en-de_2013>-cohesion=

<paragraph number_doc=3 number_total=50 number_sent=4><doc_4><rbmt-1_en-de_2013>-coherence=

<paragraph number_doc=3 number_total=50 number_sent=4><doc_4><rbmt-1_en-de_2013>-cohesion=

<paragraph number_doc=10 number_total=100 number_sent=4><doc_7><uedin_en-de_2013>-coherence=

<paragraph number_doc=10 number_total=100 number_sent=4><doc_7><uedin_en-de_2013>-cohesion=

<paragraph number_doc=5 number_total=368 number_sent=6><doc_17><rbmt-1_en-de_2013>-coherence=

<paragraph number_doc=5 number_total=368 number_sent=6><doc_17><rbmt-1_en-de_2013>-cohesion=

<paragraph number_doc=1 number_total=1 number_sent=4><doc_1><uedin_en-de_2013>-coherence=

<paragraph number_doc=1 number_total=1 number_sent=4><doc_1><uedin_en-de_2013>-cohesion=

After the student fill the form, the metadata file should look like the following:

<paragraph number_doc=2 number_total=92 number_sent=6><doc_6><rbmt-1_en-de_2013>-coherence=2

<paragraph number_doc=2 number_total=92 number_sent=6><doc_6><rbmt-1_en-de_2013>-cohesion=3

<paragraph number_doc=2 number_total=200 number_sent=4><doc_13><uedin_en-de_2013>-coherence=1
<paragraph number_doc=2 number_total=200 number_sent=4><doc_13><uedin_en-de_2013>-cohesion=3

<paragraph number_doc=3 number_total=50 number_sent=4><doc_4><rbmt-1_en-de_2013>-coherence=1
<paragraph number_doc=3 number_total=50 number_sent=4><doc_4><rbmt-1_en-de_2013>-cohesion=1

<paragraph number_doc=10 number_total=100 number_sent=4><doc_7><uedin_en-de_2013>-coherence=4
<paragraph number_doc=10 number_total=100 number_sent=4><doc_7><uedin_en-de_2013>-cohesion=4

<paragraph number_doc=5 number_total=368 number_sent=6><doc_17><rbmt-1_en-de_2013>-coherence=4
<paragraph number_doc=5 number_total=368 number_sent=6><doc_17><rbmt-1_en-de_2013>-cohesion=3

<paragraph number_doc=1 number_total=1 number_sent=4><doc_1><uedin_en-de_2013>-coherence=2
<paragraph number_doc=1 number_total=1 number_sent=4><doc_1><uedin_en-de_2013>-cohesion=1

After you finished the task, you are required to upload the metadata file with your scores in the area shared by the lecturer.

References

Akiba, Y., Sumita, E., Nakaiwa, H., Yamamoto, S., and Okuno, H. G. (2004). "Using a mixture of N-best lists from multiple MT systems in rank-sum-based confidence measure for MT outputs". In: *The 20th International Conference on Computational Linguistics*. Geneva, Switzerland, pp. 322–328.

Almaghout, H. and Specia, L. (2013). "A CCG-based Quality Estimation Metric for Statistical Machine Translation". In: *The 14th Machine Translation Summit*. Nice, France, pp. 223–230.

Avramidis, E. (2016). "Qualitative: Python tool for MT Quality Estimation supporting Server Mode and Hybrid MT". In: *The Prague Bulletin of Mathematical Linguistics* 106, pp. 147–158.

Avramidis, E., Popovic, M., Torres, D. V., and Burchardt, A. (2011). "Evaluate with confidence estimation: Machine ranking of translation outputs using grammatical features". In: *The Sixth Workshop on Statistical Machine Translation*. Edinburgh, UK, pp. 65–70.

Aziz, W., Sousa, S. C. M., and Specia, L. (2012). "PET: a tool for post-editing and assessing machine translation". In: *The 8th International Conference on Language Resources and Evaluation*. Istanbul, Turkey, pp. 3982–3987.

Aziz, W. and Specia, L. (2011). "Fully Automatic Compilation of a Portuguese-English Parallel Corpus for Statistical Machine Translation". In: *The 8th Brazilian Symposium in Information and Human Language Technology*. Cuiabá - MT, Brazil, pp. 234–238.

Bach, N., Huang, F., and Al-Onaizan, Y. (2011). "Goodness: A method for measuring machine translation confidence". In: *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, OR, pp. 211–219.

Banerjee, S. and Lavie, A. (2005). "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". In: *The ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*. Ann Harbor, MI, pp. 65–72.

Barzilay, R. and Lapata, M. (2005). "Modeling local coherence: An entity-based approach". In: *The 43rd Annual Meeting of the Association for Computational Linguistics*. Ann Harbor, MI, pp. 141–148.

Beck, D., Cohn, T., Hardmeier, C., and Specia, L. (2015). "Learning Structural Kernels for Natural Language Processing". In: *Transactions of the Association for Computational Linguistics* 3, pp. 461–473.

Ben, G., Xiong, D., Teng, Z., Lu, Y., and Liu, Q. (2013). "Bilingual Lexical Cohesion Trigger Model for Document-Level Machine Translation". In: *The 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria, pp. 382–386.

Bergstra, J. and Bengio, Y. (2012). "Random Search for Hyper-Parameter Optimization". In: *Journal of Machine Learning Research* 13, pp. 281–305.

Berka, J., Černý, M., and Bojar, O. (2011). "Quiz-based evaluation of machine translation". In: *The Prague Bulletin of Mathematical Linguistics* 95, pp. 77–86.

Biçici, E. (2016). "Referential Translation Machines for Predicting Translation Performance". In: *The First Conference on Machine Translation*. Berlin, Germany, pp. 777–781.

Biçici, E. (2013). "Referential Translation Machines for Quality Estimation". In: *The Eighth Workshop on Statistical Machine Translation*. Sofia, Bulgaria, pp. 343–351.

Biçici, E., Liu, Q., and Way, A. (2015). "Referential Translation Machines for Predicting Translation Quality and Related Statistics". In: *The Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal, pp. 304–308.

Biçici, E. and Way, A. (2014). "Referential Translation Machines for Predicting Translation Qual- ity". In: *The Ninth Workshop on Statistical Machine Translation*. Baltimore, MD, pp. 313–321.

Blain, F., Logacheva, V., and Specia, L. (2016). "Phrase Level Segmentation and Labelling of Machine Translation Errors". In: *The Tenth International Conference on Language Resources and Evaluation*. Portorož, Slovenia, pp. 2240–2245.

Blain, F., Senellart, J., Schwenk, H., Plitt, M., and Roturier, J. (2011). "Qualitative Analysis of Post-Editing for High Quality Machine Translation". In: *The 13th Machine Translation Summit*. Xiamen, China, pp. 164–171.

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., and Ueffing, N. (2004). "Confidence Estimation for Machine Translation". In: *The 20th International Conference on Computational Linguistics*. Geneva, Switzerland, pp. 315–321.

Blei, D. M. (2012). "Probabilistic Topic Models". In: *Communications of the ACM* 55.4, pp. 77–84.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). “Latent dirichlet allocation”. In: *The Journal of Machine Learning research* 3, pp. 993–1022.

Bojar, O., Graham, Y., Kamran, A., and Stanojević, M. (2016a). “Results of the WMT16 Metrics Shared Task”. In: *The First Conference on Machine Translation*. Berlin, Germany, pp. 199–231.

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2013). “Findings of the 2013 Workshop on Statistical Machine Translation”. In: *The Eighth Workshop on Statistical Machine Translation*. Sofia, Bulgaria, pp. 1–44.

Bojar, O., Buck, C., Federman, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (2014). “Findings of the 2014 Workshop on Statistical Machine Translation”. In: *The Ninth Workshop on Statistical Machine Translation*. Baltimore, MD, pp. 12–58.

Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., and Turchi, M. (2015). “Findings of the 2015 Workshop on Statistical Machine Translation”. In: *The Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal, pp. 1–46.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., and Zampieri, M. (2016b). “Findings of the 2016 Conference on Statistical Machine Translation”. In: *The First Conference on Statistical Machine Translation*. Berlin, Germany, pp. 131–198.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). “Re-evaluating the Role of BLEU in Machine Translation Research”. In: *The 11th Conference of European Chapter of the Association for Computational Linguistics*. Trento, Italy, pp. 249–256.

Callison-Burch, C., Koehn, P., Monz, C., and Schroeder, J. (2009). “Findings of the 2009 Workshop on Statistical Machine Translation”. In: *The Fourth Workshop on Statistical Machine Translation*. Athens, Greece, pp. 1–28.

Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M., and Zaidan, O. (2010). “Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation”. In: *The Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. Uppsala, Sweden, pp. 17–53.

Callison-Burch, C., Koehn, P., Monz, C., and Zaidan, O. (2011). “Findings of the 2011 Workshop on Statistical Machine Translation”. In: *The Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland, pp. 22–64.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2012). "Findings of the 2012 Workshop on Statistical Machine Translation". In: *The Seventh Workshop on Statistical Machine Translation*. Montreal, Canada, pp. 10–51.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2008). "Further Meta-Evaluation of Machine Translation". In: *The Third Workshop on Statistical Machine Translation*. Columbus, OH, pp. 70–106.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2007). "(Meta-)Evaluation of Machine Translation". In: *The Second Workshop on Statistical Machine Translation*. Prague, Czech Republic, pp. 136–158.

Carpuat, M. (2009). "One translation per discourse". In: *The Workshop on Semantic Evaluations: Recent Achievements and Future Directions*. Boulder, CO, pp. 19–27.

Carpuat, M. and Simard, M. (2012). "The Trouble with SMT Consistency". In: *The Seventh Workshop on Statistical Machine Translation*. Montreal, Quebec, Canada, pp. 442–449.

Castilho, S. and O'Brien, S. (2016). "Evaluating the Impact of Light Post-Editing on Usability". In: *The Tenth International Conference on Language Resources and Evaluation*. Portorož, Slovenia, pp. 310–316.

Castilho, S., O'Brien, S., Alves, F., and O'Brien, M. (2014). "Does post-editing increase usability? A study with Brazilian Portuguese as Target Language". In: *The 17th Annual Conference of the European Association for Machine Translation*. Dubrovnik, Croatia, pp. 183–190.

Charniak, E. (2000). "A Maximum-Entropy-Inspired Parser". In: *The 1st North American Chapter of the Association for Computational Linguistics*. Seattle, Washington, pp. 132–139.

Cortes, C. and Vapnik, V. (1995). "Support Vector Networks". In: *Machine Learning* 20.3, pp. 273–297.

David, L. (2015). "LORIA System for the WMT15 Quality Estimation Shared Task". In: *The Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal, pp. 323–329.

Day, R. R. and Park, J.-S. (2005). "Developing Reading Comprehension Questions". In: *Reading in a Foreign Language* 17.1, pp. 60–73.

Denkowski, M. and Lavie, A. (2014). "Meteor Universal: Language Specific Translation Evaluation for Any Target Language". In: *The Ninth Workshop on Statistical Machine Translation*. Baltimore, MD, pp. 376–380.

Dillinger, M. (2014). "Introduction". In: *Post-Editing of Machine Translation: Processes and Applications*. Ed. by S. O'Brien, L. W. Balling, M. Carl, and L. Specia. Cambridge Scholars Publishing.

Doherty, S. and O'Brien, S. (2014). "Assessing the Usability of Raw Machine Translated Output: A User-Centred Study using Eye Tracking". In: *International Journal of Human-Computer Interaction* 30.1, pp. 40–51.

Doherty, S. and O'Brien, S. (2009). "Can MT output be evaluated through eye tracking?" In: *The 12th Machine Translation Summit*. Ottawa, Canada, pp. 214–221.

Doherty, S., O'Brien, S., and Carl, M. (2010). "Eye Tracking as an Automatic MT Evaluation Technique". In: *Machine Translation* 24, pp. 1–13.

Eidelman, V., Boyd-Graber, J., and Resnik, P. (2012). "Topic Models of Dynamic Translation Model Adaptation". In: *The 50th Annual Meeting of the Association for Computational Linguistics*. Jeju Island, Korea, pp. 115–119.

Federico, M., Bertoldi, N., Cettolo, M., Negri, M., Turchi, M., Trombetti, M., Cattelan, A., Farina, A., Lupinetti, D., Martines, A., Massidda, A., Schwenk, H., Barrault, L., Blain, F., Koehn, P., Buck, C., and Germann, U. (2014). "The MateCat Tool". In: *The 25th International Conference on Computational Linguistics: System Demonstrations*. Dublin, Ireland, pp. 129–132.

Felice, M. and Specia, L. (2012). "Linguistic features for quality estimation". In: *The Seventh Workshop on Statistical Machine Translation*. Montreal, Quebec, Canada, pp. 96–103.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, Massachusetts: MIT Press.

Fomicheva, M. and Specia, L. (2016). "Reference Bias in Monolingual Machine Translation Evaluation". In: *54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany, pp. 77–82.

Fuji, M., Hatanaka, N., Ito, E., Kamei, S., Kumai, H., Sukehiro, T., Yoshimi, T., and Isahara, H. (2001). "Evaluation Method for Determining Groups of Users Who Find MT "Useful"". In: *The Eightth Machine Translation Summit*. Santiago de Compostela, Spain, pp. 103–108.

Fuji, M. (1999). "Evaluation Experiment for Reading Comprehension of Machine Translation Outputs". In: *The Seventh Machine Translation Summit*. Singapore, Singapore, pp. 285–289.

Gale, W. A., Church, K. W., and Yarowsky, D. (1992). "One sense per discourse". In: *The Workshop on Speech and Natural Language*. Harriman, NY, pp. 233–237.

Gandrabur, S. and Foster, G. (2003). "Confidence estimation for text prediction". In: *The Conference on Natural Language Learning*. Edmonton, Canada, pp. 95–102.

Giménez, J. and Màrquez, L. (2010). "Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation". In: *The Prague Bulletin of Mathematical Linguistics* 94, pp. 77–86.

Giménez, J. and Màrquez, L. (2009). "On the Robustness of Syntactic and Semantic Features for Automatic MT Evaluation". In: *The 4th Workshop on Statistical Machine Translation*. Athens, Greece, pp. 250–258.

Giménez, J., Màrquez, L., Comelles, E., Catellón, I., and Arranz, V. (2010). "Document-level Automatic MT Evaluation based on Discourse Representations". In: *The Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*. Uppsala, Sweden, pp. 333–338.

Gong, Z., Zhang, M., and Zhou, G. (2015). "Document-Level Machine Translation Evaluation with Gist Consistency and Text Cohesion". In: *The Second Workshop on Discourse in Machine Translation*. Lisbon, Portugal, pp. 33–40.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. (2004). "Coh-Metrix: Analysis of text on cohesion and language". In: *Behavior Research Methods, Instruments, and Computers* 36, pp. 193–202.

Graham, Y. (2015). "Improving Evaluation of Machine Translation Quality Estimation". In: *The 53rd Annual Meeting of the Association for Computational Linguistics and Seventh International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*. Beijing, China, pp. 1804–1813.

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1995). "Centering: A framework for modeling the local coherence of discourse". In: *Computational Linguistics* 21.2, pp. 203–225.

Guillou, L. (2012). "Improving Pronoun Translation for Statistical Machine Translation". In: *The Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France, pp. 1–10.

Guinaudeau, C. and Strube, M. (2013). "Graph-based Local Coherence Modeling". In: *The 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria, pp. 93–103.

Gupta, R., Orasan, C., and van Genabith, J. (2015). "ReVal: A Simple and Effective Machine Translation Evaluation Metric Based on Recurrent Neural Networks". In: *The 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal, pp. 1066–1072.

Guzmán, F., Joty, S., Màrquez, L., and Nakov, P. (2014). "Using Discourse Structure Improves Machine Translation Evaluation". In: *The 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore, MD, pp. 687–698.

Halliday, M. A. K. and Hasan, R. (1976). *Cohesion in English*. English Language Series. London, UK: Longman.

Hardmeier, C. (2014). "Discourse in Statistical Machine Translation". PhD thesis. Sweden: Department of Linguistics and Philology, Uppsala University.

Hardmeier, C. (2012). "Discourse in Statistical Machine Translation: A Survey and a Case Study". In: *Discours - Revue de linguistique, psycholinguistique et informatique* 11.

Hardmeier, C. (2011). "Improving machine translation quality prediction with syntatic tree kernels". In: *The 15th conference of the European Assoaciation for Machine Translation (EAMT 2011)*. Leuven, Belgium, pp. 233–240.

Hardmeier, C. and Federico, M. (2010). "Modelling pronominal anaphora in statistical machine translation". In: *The 7th International Workshop on Spoken Language Translation*. Paris, France, pp. 283–289.

Hardmeier, C., Nivre, J., and Tiedemann, J. (2012). "Tree Kernels for Machine Translation Quality Estimation". In: *The Seventh Workshop on Statistical Machine Translation*. Montréal, Canada, pp. 109–113.

Hardmeier, C., Nakov, P., Stymne, S., Tiedemann, J., Versley, Y., and Cettolo, M. (2015). "Pronoun-Focused MT and Cross-Lingual Pronoun Prediction: Findings of the 2015 DiscoMT Shared Task on Pronoun Translation". In: *The Second Workshop on Discourse in Machine Translation*. Lisbon, Portugal, pp. 1–16.

He, Y., Ma, Y., van Genabith, J., and Way, A. (2010). "Bridging SMT and TM with Translation Recommendation". In: *The 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden, pp. 622–630.

Hobbs, J. R. (1979). "Coherence and Coreference". In: *Cognitive Science* 3, pp. 67–90.

Jones, D. A., Shen, W., Granoien, N., Herzog, M., and Weinstein, C. (2005a). "Measuring Human Readability of Machine Generated Text: Studies in Speech Recognition and Machine Translation". In: *The IEEE International Conference on Acoustics, Speech, and Signal Processing*. Philadelphia, PA.

Jones, D. A., Gibson, E., Shen, W., Granoien, N., Herzog, M., Reynolds, D., and Weinstein, C. (2005b). "Measuring Translation Quality by Testing English Speakers with a New Defense Language Proficiency Test for Arabic". In: *The International Conference on Intelligence Analysis*. McLean, VA.

Joty, S., Carenini, G., Ng, R. T., and Mehdad, Y. (2013). "Combining Intra- and Multi-sentential Rhetorical Parsing for Document-level Discourse Analysis". In: *The 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria, pp. 486–496.

Joty, S., Guzmán, F., Màrquez, L., and Nakov, P. (2014). "DiscoTK: Using discourse structure for machine translation evaluation". In: *The Ninth Workshop on Statistical Machine Translation*. Baltimore, MD, pp. 403–408.

Kenter, T., Borisov, A., and Rijke, M. d. (2016). "Siamese CBOW: Optimizing Word Embeddings for Sentence Representations". In: *54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany, pp. 941–951.

Klerke, S., Castilho, S., Barrett, M., and Søgaard, A. (2015). "Reading metrics for estimating task efficiency with MT output". In: *The Sixth Workshop on Cognitive Aspects of Computational Language Learning*. Lisbon, Portugal, pp. 6–13.

Koehn, P. (2010). *Statistical Machine Translation*. Marina del Rey, CA: Information Science Institute.

Koehn, P. and Monz, C. (2006). "Manual and Automatic Evaluation of Machine Translation between European Languages". In: *The Workshop on Statistical Machine Translation*. New York City, NY, pp. 102–121.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). "MOSES: Open source Toolkit for Statistical Machine Translation". In: *The Annual Meeting of the Association for Computational Linguistics, demonstration session*. Prague, Czech Republic, pp. 177–180.

Koponen, M., Aziz, W., Ramos, L., and Specia, L. (2012). "Post-editing time as a measure of cognitive effort". In: *The AMTA 2012 Workshop on Post-Editing Technology and Practice*. San Diego, CA, pp. 11–20.

Kozlova, A., Shmatova, M., and Frolov, A. (2016). "YSDA Participation in the WMT'16 Quality Estimation Shared Task". In: *The First Conference on Machine Translation*. Berlin, Germany, pp. 793–799.

Krings, H. P. (2001). *Repairing texts: Empirical investigations of machine translation post-editing process*. Kent, OH: The Kent State University Press.

Lacruz, I., Denkowski, M., and Lavie, A. (2014). "Cognitive Demand and Cognitive Effort in Post-Editing". In: *The Third Workshop on Post-Editing Technology and Practice*. Vancouver, Canada, pp. 73–84.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). "An Introduction to Latent Semantic Analysis". In: *Discourse Processes* 25, pp. 259–284.

Le, Q. and Mikolov, T. (2014). "Distributed Representations of Sentences and Documents". In: *The 31st International Conference on Machine Learning*. Beijing, China, pp. 1188–1196.

LeNagard, R. and Koehn, P. (2010). “Aiding Pronoun Translation with Co-Reference Resolution”. In: *The Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*. Uppasala, Sweden, pp. 252–261.

Li, J. J., Carpuat, M., and Nenkova, A. (2014). “Assessing the Discourse Factors that Influence the Quality of Machine Translation”. In: *The 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore, MD, pp. 283–288.

Lin, R., Liu, S., Yang, M., Li, M., Zhou, M., and Li, S. (2015). “Hierarchical Recurrent Neural Network for Document Modeling”. In: *The 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal, pp. 899–907.

Logacheva, V. and Specia, L. (2015). “Phrase-level Quality Estimation for Machine Translation”. In: *The 12th International Workshop on Spoken Language Translation*. Da Nang, Vietnam, pp. 143–150.

Louis, A. and Nenkova, A. (2012). “A coherence model based on syntactic patterns”. In: *The 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island, Korea, pp. 1157–1168.

Louis, A. and Nenkova, A. (2013). “Automatically Assessing Machine Summary Content Without a Gold Standard”. In: *Computational Linguistics* 39.2, pp. 267–300.

Luong, N.-Q. (2014). “Word Confidence Estimation for Statistical Machine Translation”. PhD thesis. Grenoble, France: Laboratoire d’Informatique de Grenoble.

Macháček, M. and Bojar, O. (2013). “Results of the WMT13 Metrics Shared Task”. In: *The Eighth Workshop on Statistical Machine Translation*. Sofia, Bulgaria, pp. 45–51.

Macháček, M. and Bojar, O. (2014). “Results of the WMT14 Metrics Shared Task”. In: *The Ninth Workshop on Statistical Machine Translation*. Baltimore, MD, pp. 293–301.

Mann, W. C. and Thompson, S. A. (1987). *Rhetorical Structure Theory: A Theory of Text Organization*. Cambridge, UK: Cambridge University Press.

Marcu, D., Carlson, L., and Watanabe, M. (2000). “The automatic translation of discourse structures”. In: *The 1st North American chapter of the Association for Computational Linguistics conference*. Seattle, WA, pp. 9–17.

Martínez-Garcia, E., Bonet, C. España, and Màrquez, L. (2015). “Document-level machine translation with word vector models”. In: *The 18th Annual Conference of the European Association for Machine Translation*. Antalya, Turkey, pp. 59–66.

Martínez-Garcia, E., Bonet, C. España, Tiedemann, J., and Màrquez, L. (2014). “Word’s Vector Representations meet Machine Translation”. In: *The Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Doha, Qatar, pp. 132–134.

Martins, A. F. T., Almeida, M. B., and Smith, N. A. (2013). "Turning on the Turbo: Fast Third-Order Non-Projective Turbo Parsers". In: *The 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria, pp. 617–622.

Martins, A. F. T., Astudillo, R., Hokamp, C., and Kepler, F. N. (2016). "Unbabel's Participation in the WMT16 Word-Level Translation Quality Estimation Shared Task". In: *The First Conference on Machine Translation*. Berlin, Germany, pp. 806–811.

Meurers, R. Z., Ott, N., and Kopp, J. (2011). "Evaluating Answers to Reading Comprehension Questions in Context: Results for German and the Role of Information Structure". In: *The TextInfer 2011 Workshop on Textual Entailment*. Edinburgh, UK, pp. 1–9.

Meyer, T. and Popescu-Belis, A. (2012). "Using sense-labeled discourse connectives for statistical machine translation". In: *The Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*. Avignon, France, pp. 129–138.

Meyer, T., Popescu-Belis, A., Hajlaoui, N., and Gesmundo, A. (2012). "Machine translation of labeled discourse connectives". In: *The Tenth Biennial Conference of the Association for Machine Translation in the Americas*. San Diego, CA.

Meyer, T., Popescu-belis, A., Zufferey, S., and Cartoni, B. (2011). "Multilingual Annotation and Disambiguation of Discourse Connectives for Machine Translation". In: *The 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Portland, OR, pp. 194–203.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013). "Exploiting Similarities among Languages for Machine Translation". In: *CoRR* abs/1309.4168. URL: http://arxiv.org/abs/1309.4168.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013a). "Distributed representations of words and phrases and their compositionality". In: *The 26th International Conference on Neural Information Processing Systems*. Lake Tahoe, Nevada, pp. 3111–3119.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013b). "Efficient Estimation of Word Representations in Vector Space". In: *The International Conference on Learning Representations 2013: workshop track*. Scottsdale, AZ, pp. 1–12.

Mitkov, R. (2004). "Anaphora Resolution". In: *The Oxford Handbook of Computational Linguistics*. Ed. by R. Mitkov. Oxford University Press.

Moschitti, A. (2006). "Making tree kernels practical for natural language learning". In: *The Eleventh International Conference on European Association for Computational Linguistics*. Trento, Italy, pp. 113–120.

Nirenburg, S. (1993). *Progress in Machine Translation*. Amsterdam, Netherlands: IOS B. V.

Novák, M., Oele, D., and van Noord, G. (2015). "Comparison of Coreference Resolvers for Deep Syntax Translation". In: *The Second Workshop on Discourse in Machine Translation*. Lisbon, Portugal, pp. 17–23.

O'Brien, S. (2011). "Towards predicting post-editing productivity". In: *Machine Translation* 25, pp. 197–215.

Ott, N., Ziai, R., and Meurers, D. (2012). "Creation and analysis of a reading comprehension exercise corpus: Towards evaluating meaning in context". In: *Multilingual Corpora and Multilingual Corpus Analysis*. Ed. by T. Schmidt and K. Worner. Hamburg Studies on Multilingualism (Book 14). Amsterdam, The Netherlands: John Benjamins Publishing Company, pp. 47–69.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). "The Proposition Bank: A Corpus Annotated with Semantic Roles". In: *Computational Linguistics* 31.1, pp. 72–105.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. jing (2002). "BLEU: a Method for Automatic Evaluation of Machine Translation". In: *The 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, PA, pp. 311–318.

Parra Escartín, C. and Arcedillo, M. (2015). "Machine translation evaluation made fuzzier: A study on post-editing productivity and evaluation metrics in commercial settings". In: *The Fourth Workshop on Hybrid Approaches to Translation*. Beijing, China, pp. 40–45.

Pedregosa-Izquierdo, F. (2015). "Feature extraction and supervised learning on fMRI : from practice to theory". PhD thesis. Paris, France: École Doctorale Informatique, Télécommunications et Électronique, Université Pierre et Marie Curie.

Pighin, D and Màrquez, L (2011). "Automatic projection of semantic structures: an application to pairwise translation ranking". In: *The SSST-5, Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Portland, OR, pp. 1–9.

Pitler, E. and Nenkova, A. (2009). "Using Syntax to Disambiguate Explicit Discourse Connectives in Text". In: *The Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the Fourth International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*. Suntec, Singapore, pp. 13–16.

Plitt, M. and Masselot, F. (2010). "A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context". In: *The Prague Bulletin of Mathematical Linguistics* 93, pp. 7–16.

Potet, M., Esperança-Rodier, E., Besacier, L., and Blanchon, H. (2012). "Collection of a Large Database of French-English SMT Output Corrections". In: *The 8th International Conference on Language Resources and Evaluation*. Istanbul, Turkey, pp. 23–25.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). "The Penn Discourse Treebank 2.0". In: *The 6th International Conference on Language Resources and Evaluation*. Marrackech, Morocco, pp. 2961–2968.

Quirk, C. B. (2004). "Training a sentence-level machine translation confidence metric". In: *The International Conference on Language Resources and Evaluation*. Lisbon, Portugal, pp. 825–828.

Ramsay, A. (2004). "Discourse". In: *The Oxford Handbook of Computational Linguistics*. Ed. by R. Mitkov. Oxford University Press.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. Cambridge, Massachusetts: MIT Press.

Řehůřek, R. and Sojka, P. (2010). "Software Framework for Topic Modelling with Large Corpora". In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta, pp. 45–50.

Rennie, J. D. M. and Srebro, N. (2005). "Loss functions for preference levels: Regression with discrete ordered labels". In: *The IJCAI Multidisciplinary Workshop on Advances in Preference Handling*. Edinburgh, Scotland, pp. 180–186.

Richardson, M., Burges, C. J. C., and Renshaw, E. (2013). "MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text". In: *The 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, WA, pp. 193–203.

Rogers, S. and Girolami, M. (2012). *A first course in Machine Learning*. Boca Raton, FL: CRC Press, Taylor and Francis Group.

Rubino, R., Souza, J. G. C. d., Foster, J., and Specia, L. (2013). "Topic Models for Translation Quality Estimation for Gisting Purposes". In: *The 14th Machine Translation Summit*. Nice, France, pp. 295–302.

Sagemo, O. and Stymne, S. (2016). "The UU Submission to the Machine Translation Quality Estimation Task". In: *The First Conference on Machine Translation*. Berlin, Germany, pp. 825–830.

Sajjad, H., Guzman, F., Durrani, N., Bouamor, H., Abdelali, A., Temnikova, I., and Vogel, S. (2016). "Eyes Don't Lie: Predicting Machine Translation Quality Using Eye Movement". In: *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, CA, pp. 1082–1088.

Scarton, C. (2015). "Discourse and Document-level Information for Evaluating Language Output Tasks". In: *The 2015 Conference of the North American Chapter of the Association*

for Computational Linguistics: Student Research Workshop. Denver, Colorado, pp. 118–125.

Scarton, C. and Specia, L. (2015). "A quantitative analysis of discourse phenomena in machine translation". In: *Discours - Revue de linguistique, psycholinguistique et informatique* 16.

Scarton, C. and Specia, L. (2016). "A Reading Comprehension Corpus for Machine Translation Evaluation". In: *The Tenth International Conference on Language Resources and Evaluation*. Portorož, Slovenia, pp. 3652–3658.

Scarton, C. and Specia, L. (2014a). "Document-level translation quality estimation: exploring discourse and pseudo-references". In: *The 17th Annual Conference of the European Association for Machine Translation*. Dubrovnik, Croatia, pp. 101–108.

Scarton, C. and Specia, L. (2014b). "Exploring Consensus in Machine Translation for Quality Estimation". In: *The Ninth Workshop on Statistical Machine Translation*. Baltimore, Maryland, USA, pp. 342–347.

Scarton, C., Tan, L., and Specia, L. (2015). "USHEF and USAAR-USHEF participation in the WMT15 QE shared task". In: *The Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal, pp. 336–341.

Scarton, C., Zampieri, M., Vela, M., van Genabith, J., and Specia, L. (2015). "Searching for Context: a Study on Document-Level Labels for Translation Quality Estimation". In: *The 18th Annual Conference of the European Association for Machine Translation*. Antalya, Turkey, pp. 121–128.

Scarton, C., Beck, D., Shah, K., Sim Smith, K., and Specia, L. (2016). "Word embeddings and discourse information for Quality Estimation". In: *The First Conference on Machine Translation*. Berlin, Germany, pp. 831–837.

Schmid, H. (1994). "Probabilistic Part-of-Speech Tagging Using Decision Trees". In: *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK.

Servan, C., Le, N.-T., Luong, N. Q., Lecouteux, B., and Besacier, L. (2011). "Evaluating Answers to Reading Comprehension Questions in Context: Results for German and the Role of Information Structure". In: *The TextInfer 2011 Workshop on Textual Entailment*. Edinburgh, UK, pp. 1–9.

Shah, K., Cohn, T., and Specia, L. (2013). "An Investigation on the Effectiveness of Features for Translation Quality Estimation". In: *The 14th Machine Translation Summit*. Nice, France, pp. 167–174.

Shah, K., Ng, R. W. M., Bougares, F., and Specia, L. (2015a). "Investigating Continuous Space Language Models for Machine Translation Quality Estimation". In: *The 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal, pp. 1073–1078.

Shah, K., Logacheva, V., Paetzold, G. H., Blain, F., Beck, D., Bougares, F., and Specia, L. (2015b). "SHEF-NN: Translation Quality Estimation with Neural Networks". In: *The Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal, pp. 342–347.

Sim Smith, K., Aziz, W., and Specia, L. (2016a). "Cohere: A Toolkit for Local Coherence". In: *The Tenth International Conference on Language Resources and Evaluation*. Portorož, Slovenia, pp. 4111–4114.

Sim Smith, K., Aziz, W., and Specia, L. (2016b). "The Trouble With Machine Translation Coherence". In: *The 19th Annual Conference of the European Association for Machine Translation*. Riga, Latvia, pp. 178–189.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). "A Study of Translation Edit Rate with Targeted Human Annotation". In: *The Seventh biennial conference of the Association for Machine Translation in the Americas*. Cambridge, MA, pp. 223–231.

Snover, M., Madnani, N., Dorr, B. J., and Schwartz, R. (2009). "Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric". In: *The Fourth Workshop on Statistical Machine Translation*. Athens, Greece, pp. 259–268.

Soricut, R., Bach, N., and Wang, Z. (2012). "The SDL Language Weaver Systems in the WMT12 Quality Estimation Shared Task". In: *The Seventh Workshop on Statistical Machine Translation*. Montréal, Canada, pp. 145–151.

Soricut, R. and Echihabi, A. (2010). "TrustRank: Inducing Trust in Automatic Translations via Ranking". In: *The 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden, pp. 612–621.

Soricut, R. and Marcu, D. (2003). "Sentence Level Discourse Parsing using Syntactic and Lexical Information". In: *The 2003 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Edmonton, Canada, pp. 149–156.

Soricut, R. and Narsale, S. (2012). "Combining Quality Prediction and System Selection for Improved Automatic Translation Output". In: *The Seventh Workshop on Statistical Machine Translation*. Montréal, Canada, pp. 163–170.

Specia, L. and Farzindar, A. (2010). "Estimating machine translation post-editing effort with HTER". In: *Proceedings of AMTA-2010 Workshop Bringing MT to the User: MT Research and the Translation Industry*. Denver, CO, pp. 33–41.

Specia, L., Paetzold, G., and Scarton, C. (2015). "Multi-level Translation Quality Prediction with QuEst++". In: *The 53rd Annual Meeting of the Association for Computational Linguistics and Seventh International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing: System Demonstrations*. Beijing, China, pp. 115–120.

Specia, L., Turchi, M., Cancedda, N., Dymetman, M., and Cristianini, N. (2009a). "Estimating the Sentence-Level Quality of Machine Translation Systems". In: *The 13th Annual Conference of the European Association for Machine Translation*. Barcelona, Spain, pp. 28–37.

Specia, L., Turchi, M., Wang, Z., Shawe-Taylor, J., and Saunders, C. (2009b). "Improving the Confidence of Machine Translation Quality Estimates". In: *The 12th Machine Translation Summit*. Ottawa, Ontario, Canada, pp. 136–143.

Specia, L., Hajlaoui, N., Hallet, C., and Aziz, W. (2011). "Predicting machine translation adequacy". In: *The 13th Machine Translation Summit*. Xiamen, China, pp. 19–23.

Specia, L., Shah, K., Souza, J. G. de, and Cohn, T. (2013). "QuEst - A translation quality estimation framework". In: *The 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Sofia, Bulgaria, pp. 79–84.

Stanojević, M., Kamran, A., Koehn, P., and Bojar, O. (2015). "Findings of the 2015 Workshop on Statistical Machine Translation". In: *The Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal, pp. 1–46.

Stede, M. (2011). *Discourse Processing*. en. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Steele, D. and Specia, L. (2016). "Predicting and Using Implicit Discourse Elements in Chinese-English Translation". In: *Baltic J. Modern Computing* 4.2, pp. 305–317.

Stymne, S., Danielsson, H., Bremin, S., Hu, H., Karlsson, J., Lillkull, A. P., and Wester, M. (2012). "Eye Tracking as a Tool for Machine Translation Error Analysis". In: *The 8th International Conference on Language Resources and Evaluation*. Istanbul, Turkey, pp. 1121–1126.

Tomita, M., Masako, S., Tsutsumi, J., Matsumura, M., and Yoshikawa, Y. (1993). "Evaluation of MT Systems by TOEFL". In: *The Fifth International Conference on Theoretical and Methodological Issues in Machine Translation*. Kyoto, Japan, pp. 252–265.

Tu, M., Zhou, Y., and Zong, C. (2013). "A Novel Translation Framework Based on Rhetorical Structure Theory". In: *The 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria, pp. 370–374.

Turchi, M., Steinberger, J., and Specia, L. (2012). "Relevance Ranking for Translated Texts". In: *The 16th Annual Conference of the European Association for Machine Translation*. Trento, Italy, pp. 153–160.

Ture, F., Oard, D. W., and Resnik, P. (2012). "Encouraging consistent translation choices". In: *The 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Montreal, Canada, pp. 417–426.

Turian, J. P., Shen, L., and Melamed, I. D. (2003). "Evaluation of Machine Translation and its Evaluation". In: *The Ninth Machine Translation Summit*. New Orleans, LA, pp. 386–393.

Ueffing, N. (2006). "Word Confidence Measures for Machine Translation". PhD thesis. Aachen, Germany: Computer Science Department, RWTH Aachen University.

Ueffing, N., Macherey, K., and Ney, H. (2003). "Confidence measures for statistical machine translation". In: *The Ninth Machine Translation Summit*. New Orleans, LA, pp. 394–401.

Ueffing, N. and Ney, H. (2007). "Word-Level Confidence Estimation for Machine Translation". In: *Computational Linguistics* 33.1, pp. 1–40.

Ueffing, N. and Ney, H. (2005). "Word-level confidence estimation for machine translation using phrase-based translation models". In: *The Human Language Technology Conference*. Vancouver, Canada, pp. 763–770.

Wisniewski, G., Singh, A. K., Segal, N., and Yvon, F. (2013). "Design and Analysis of a Large Corpus of Post-Edited Translations: Quality Estimation, Failure Analysis and the Variability of Post-Edition". In: *The 14th Machine Translation Summit*. Nice, France, pp. 117–124.

Wong, B. T. M. and Kit, C. (2012). "Extending machine translation evaluation metrics with lexical cohesion to document level". In: *The 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island, Korea, pp. 1060–1068.

Xiao, T., Zhu, J., Yao, S., and Zhang, H. (2011). "Document-level consistency verification in machine translation". In: *The 13th Machine Translation Summit*. Xiamen, China, pp. 131–138.

Xiong, D., Ben Guosheng Zhang, M., and Liu, Q. (2011). "Modeling Lexical Cohesion for Document-Level Machine Translation". In: *The 23rd international joint conference on Artificial Intelligence*. Beijing, China, pp. 2183–2189.

Xiong, D., Zhang, M., and Li, H. (2010). "Error detection of statistical machine translation using linguistic features". In: *The 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden, pp. 604–611.

Zhao, B. and Zing, E. (2008). "HM-BiTAM: bilingual topic exploration, word alignment and translation". In: *Advances in Neural Information Processing Systems*. Ed. by J. C. Platt, D. Koller, Y. Singer, and S. Rowels. 20th ed. Cambridge, MA: MIT Press, pp. 1689–1696.

Zhengxian, G., Yu, Z., and Guodong, Z. (2010). "Statistical Machine Translation Based on LDA". In: *The Fourth International Universal Communication Symposium*. Beijing, China, pp. 279–283.