The University Of Sheffield.

Discourse Cohesion in Chinese-English Statistical Machine Translation

By:

David Steele

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Sheffield
Department of Computer Science
Sheffield, UK

September, 2019

Acknowledgements

First, I should like to thank my two boys, Jaq and Nico, without whose excellent behaviour and attitudes throughout my studies and full time working, I should not have been able to get this far. Also thanks go to my wife, Wanlin, for her kind support in the early stages of my work and her guidance on the finer aspects of the Chinese language.

I should also like to thank Mark Hepple and Roger Moore on my panel for their helpful suggestions.

Also thanks to my colleagues in the NLP group who have always filled the labs with so much energy, and created a positive vibe.

I also need to thank my parents for their kind and supportive words throughout my studies, with an extra mention to my mum who accompanied me on a trip to the US to act as a babysitter for my two boys whilst I attended the NAACL conference.

Finally, I should very much like to express my gratitude to Lucia for having offered me the opportunity to undertake the PhD, and for her ongoing support during my time at Sheffield. Lucia always maintained a high level of professionalism throughout my studies, and she also offered keen insights into so many facets of her knowledge domain. In addition, she showed an enduring level of patience when it came to my meeting deadlines and when trying to move forward, for all of which I shall always be grateful.

Abstract

In discourse, cohesion is a required component of meaningful and well organised text. It establishes the relationship between different elements in the text using a number of devices such as pronouns, determiners, and conjunctions.

In translation a well translated document will display the correct cohesion and use of cohesive devices that are pertinent to the language. However, not all languages have the same cohesive devices or use them in the same way. In statistical machine translation this is a particular barrier to generating smooth translations, especially when sentences in parallel corpora are being treated in isolation and no extra meaning or cohesive context is provided beyond the sentential level.

In this thesis, focussing on Chinese[1] and English as the language pair, we examine discourse cohesion in statistical machine translation looking at ways that systems can leverage discourse cues and signals in order to produce smoother translations. We also provide a statistical model that improves translation output by adding additional tokens within text that can be used to leverage extra information.

A significant part of this research involved visualising many of the results and system outputs, and so an overview of two important pieces of visualisation software that we developed is also included.

[1] For this thesis Chinese means the primary language spoken in mainland China.

"The original is unfaithful to the translation"

–Jorge Luis Borges

Contents

List of Figures

List of Tables

List of Acronyms

BEER BEtter Evaluation as Ranking

BLEU Bilingual Evaluation Understudy

DARPA Defense Advanced Research Projects Agency

DCs Discourse Connectives

DMs Discourse Markers

EBMT Example Based Machine Translation

E-Dist Edit Distance

GALE Global Autonomous Language Exploitation

GPU Graphics Processing Unit

GPUs Graphics Processing Units

GUI Graphical User Interface

HPBT Hierarchical Phrase Based Translation

HMM Hidden Markov Model

iBleu Interactive BLEU

LSTM Long Short-Term Memory (Neural Network)

METEOR Metric for Evaluation of Translation with Explicit ORdering

MV Metric Viewer

MT Machine Translation

NER Named Entity Recognition

NIST National Institute of Standards and Technology

NLP Natural Language Processing

NMT Neural Machine Translation

PBT Phrase Based Translation

POS Part-of-Speech

PNG Portable Network Graphics

RBMT Rule Based Machine Translation

RNN Recurrent Neural Network

SBT Syntax Based Translation

SMT Statistical Machine Translation

TER Translation Edit Rate

TBT Tree Based Translation

VEMV Vis-Eval Metric Viewer

VM VisMet

WA Word Alignment

WAs Word Alignments

WBT Word Based Translation

WER Word Error Rate

WMT Workshop on Statistical Machine Translation

Chapter 1

1 Introduction

With the world becoming more globalised and interconnected there is an ever increasing need for translation and interpreting services. Demands for these services are high in many areas such as politics, business, security, and electronics, to name a few. To meet these demands there is a reliance on Machine Translation (MT).

With vast improvements being made to MT in recent years it has become an integral part of translation services, helping translators to improve their performance through better quality automated translations that require fewer post-processing steps.

This research comes at a time when progress in MT has been significant and there has been a gradual move away from Statistical Machine Translation (SMT) (the main focus of this thesis) and toward Neural Machine Translation (NMT). At the time of writing SMT was still a well used paradigm, but NMT was gaining traction as the technology improved and access to the requisite Graphics Processing Units (GPUs) became easier.

However, despite the improvements in SMT and new technologies being embraced, translation quality is still often far from perfect. This is due to many reasons such as: corpus quality and availability, lack of context or small context windows with sentences and words commonly being treated in isolation, and divergence between language usage, to name just a few. One particular problem with respect to MT, and the target of this thesis, is around discourse cohesion (e.g. discourse relations and discourse connectives).

A more in depth review of grammar related to cohesion of text is provided in Chapter 3. However, in Section 1.1 we outline the underlying problems that the usage and spread

of cohesive devices across languages can cause for leading SMT systems.

1.1 The Problem

When learning a new language, learners often have first language interference, where, for example, the grammatical or part-of-speech (POS) differences between the native language or 'mother tongue' and the new language cause confusion or 'interference' for the speaker.

When it comes to MT these differences between languages are a particular problem. For example, languages such as Chinese and Japanese have no article systems and cause a great deal of difficulty when being translated into English (Swan and Smith, 2004).

Despite substantial improvements in both SMT (and NMT), which have enhanced the accuracy of automated translations, MT systems are still unable to deliver human quality translations in many cases. The problem is especially prominent with complex composite sentences and distant language pairs.

Rather than considering larger discourse segments as a whole, SMT systems in particular focus on the translation of single sentences independently, and clauses tend to be treated in isolation. This leads to a loss of contextual information and little or no recognition of inference.

However, texts and discourse in natural language do not normally consist of isolated chunks. They are more commonly made up of connected sentences, which in turn contain clauses within them to create a unified whole; in other words a collection of meaningful sentences, which have a particular communicative purpose. Here the reference to 'connected' and 'meaningful' highlights that serious discourse is expected to be both cohesive and coherent (Clark et al., 2013).

Discourse Markers (DMs), for example, are viewed as essential contextual links between the various discourse segments. Despite the fundamental role DMs have in terms of lexical cohesion, SMT systems often do not explicitly address DM constructions and the way they work. This therefore can result is translations that often lack the cohesive cues otherwise provided in normal texts.

In addition, elements such as DMs are routinely translated into the target language in a myriad of ways that differ from how they are used in the source language (Hardmeier, 2012; Meyer and Webber, 2013). Furthermore, it has been shown that single DMs can actually signal a variety of discourse relations depending on their location and frequency and current SMT systems are unable to adequately recognise or distinguish between each during the translation process (Hajlaoui and Popescu-Belis, 2013).

So far this discussion has assumed that the devices are usually explicit across languages, where this is indeed not the case. Both English and Chinese make use of explicit and implicit markers in many situations, with the much greater challenge coming from tackling implicit DMs, which in turn leads to implicit discourse relations. In Chinese especially, there is an abundance of implicit discourse relations, much more than in English, that occur both within sentences as well as across wider discourse segments (Yung, 2014). A typical Chinese sentence is given in Example 1, along with a literal translation and a given corpus (BTEC) translation:

(1) 如果考尔曼[2] 先生不在, 我可以见其他负责人[吗] ? (Chinese)

if Coleman mr not in, I can see another responsible person [question particle]?

(Literal)

if mr. coleman is out , then may I see another person in charge ?

(BTEC translation)

Comparing the literal translation with the given translation, it is clear that the 'then' condition is implicit in the Chinese sentence, but has been made explicit in the English. Inputting the Chinese sentence into Google Translate, produces the following translation:

(2) *'Mr. Colman[3] if not, I can see the other person in charge of it?'*

(Google - SMT Version - 2016)

[2]The three characters 考尔曼are pronounced as kao-er-man (written in pinyin), a phonetic representation of Coleman/Colman.

[3]Google spells Coleman as Colman in this case.

The output from Google remains fairly close to the given literal translation (even with the name), but is nowhere near as smooth as the corpus translation. It completely misses the 'if - then' relation.

This shows that even at the sentence level there exists a local context, which produces dependencies between certain words (e.g links between clauses). The cohesion information within the sentence can hold vital clues and so it is important to try to capture this information. As such we contend that this 'contextual' information, in the very least, could be used to guide translations in order to improve accuracy.

Another example (see Chapter 5, Example 11) is given in Example 3:

(3) 他因为病了,没来上课。

The literal translation of this sentence is: *he because ill, not come class*.

From the literal translation it can be seen that the pronoun for he (他) is only needed once in the Chinese, and after that, since the person has been established, there is enough information from the context to have a full and complete meaningful sentence.

However, a typical English translation could be:

(4) *because he was ill, he did not come to class*.

In this case a second 'he' is required to make the sentence more fluid. For an SMT system this relation can be difficult to spot as (amongst other things) each of the two segments in the sentence (separated by the comma) may be treated in isolation. This is discussed further in Chapter 5.

One further example as motivation is a short sentence (in Chinese) that actually carries so much information and context:

(5) 她一学就会。

A literal translation for Example 5 is: *she one study then can*.

However, a literal translation loses so much of the function of this sentence as, in this case, the characters 一 and 就 form part of a wider grammatical construct, which for brevity we shall state is a 'as soon as' relationship. That is, 'as soon as' something happens then something else happens as a result.

A good translation of Example 5 could be:

(6) *as soon as she studies it then she can do it.*

In this English translation given in Example 6 there is a lot of information that has to be extracted from the delicate relation of the five Chinese characters in order to generate a smooth sentence. With SMT systems many existing models appear to focus on producing well-translated localised sentence fragments, but often ignore the wider cohesion and contextual clues or devices that fall within the sentence. While some developments in SMT potentially allow the modelling of discourse information (Hardmeier et al., 2013), limited resources have been dedicated to addressing many of the devices used. Despite this, there has been some work towards including cohesion models, be it using lexical chains or enhancing grammatical cohesion. It has been suggested that annotated corpora, which hold discourse-connective information can be utilised for predicting implicitation and eventually used to guide a discourse-aware SMT system (Yung, 2014).

In this work we explore some of the above ideas, and following a corpus analysis, propose our own model for improving translation through marking discourse relations in sentences.

1.2 Advances Through Neural Machine Translations

The absolute focus of this work is on SMT, but since, at the time of writing, there has been a strong shift away from SMT and toward NMT it is important to note that NMT does partially solve some of the expressed problems, although some issues do still remain.

Both Google Translate and Microsoft Bing Translator, two of the leading commercial translation systems, now employ neural network technology for MT. It shows much improved results over previous iterations and partially solves some of the problems we have explored.

For example,

(7) 他因为病了,没来上课。

is now translated as:

He didn't come to class because he was ill. (Bing).

Previously using Bing the Chinese sentence in Example 7 was translated as:

he is ill, absent.

Clearly, the new version of Bing is much better, and can handle the relations in the sentence. However, some sentences still offer a challenge, especially where a lot of information is being inferred:

(8) 她一学就会。

This is translated as:

She will when she learns. (Bing)

She will learn as soon as she learns. (Google)

It appears that there is still a lot of information that needs to be extracted from the sentence. With the translation from Google, despite the output being semantically meaningless and sounding somewhat snippy it does seem to be trying to address the 'as soon as' construct.

One final example is given here in Example 9 to highlight how the problems that we are looking at for SMT still remain on some level for NMT as well.

(9) 怎么一吃火锅就拉肚子?

The literal translation for this sentence is: *how once eat hot pot then diarrhea?*

A smooth translation of this sentence could be: *Why do you have diarrhea when you eat hot pot?*

The output from Google (17/07/19) is: *How do you eat a hot pot and diarrhea?*

Whilst most of the information is captured in the translation, the relation has not been fully realised and leads to a somewhat awkward sentence.

It is clear that recent development in MT, more specifically a move toward NMT, addresses some of the problems we highlight in this work (but not for SMT). However, many issues still exist that remain unresolved and we contend that the explicitation model or method we propose for SMT could actually be used on some level to benefit NMT as well.

1.3 Scope and Aims

While MT has been improving and is becoming more widely used, the quality is often still lacking, and there is an increasing awareness of the need to integrate more linguistic information, including, for the purposes of this thesis, cohesion. Modelling cohesion in MT is a difficult task. To that end, the main aims of our work can be summarised with the following two questions:

1. What are the limitations of current strong SMT approaches (e.g. hierarchical tree-based) in terms of handling cohesive devices?

2. How can we better model cohesive devices within sentences?

SMT is the primary focus of this work and it has to be noted that, to date, we have not worked extensively on NMT models save for some brief observations made along the way. The content of this thesis is therefore purely aimed at making improvements in SMT. Whilst we have not explored NMT in depth, we still believe that the explicitation approach that we describe should also work for that as well.

1.4 Contributions

Here we provide an outline of the contributions of this work, which largely feature in Chapters 5, 6, and 7, along with the literature review in Chapter 4. A full list of our publications is included in the Appendix.

1. Divergences in the Usage of Discourse Markers in English and Mandarin Chinese (Steele and Specia, 2014). This, through examining divergence, highlights a number of structural differences in composite sentences from parallel corpora. It shows examples of how SMT systems deal with the differences. (Published in TSD, 2014)

2. Improving the Translation of Discourse Markers for Chinese into English (Steele, 2015). This focusses on the difficulties of dealing with DMs in SMT and looks at initial ways to model DMs within sentences. (Published in NAACL-HLT (Student Research Workshop), 2015)

3. Predicting and Using Implicit Discourse Elements in Chinese-English Translation (Steele and Specia, 2016) . This introduces a prediction model used to insert tokens into a source language, which then can in turn be leveraged in order to give better translations by providing extra information for the MT systems to work with. (Published in EAMT, 2016)

4. WA-Continuum: Visualising Word Alignments across Multiple Parallel Sentences Simultaneously (Steele and Specia, 2015). This describes a tool we built that was created specifically with the purpose of visualising word alignments for SMT in great detail. It was an essential tool for our work as we needed to look at the changes to numerous word alignments in a robust and quick fashion, and in a way that could be visualised far more easily then with text alone. (Published in ACL-IJCNLP, (System Demonstrations), 2015)

5. Vis-Eval Metric Viewer: A Visualisation Tool for Inspecting and Evaluating Metric Scores of Machine Translation Output (Steele and Specia, 2018). Like the WA-Continuum tool, this was required for our research. With each change to our prediction model we had to examine various MT system outputs in a speedy and robust manner. We needed a tool that could produce an array of meaningful metric scores for our translations and that could search the voluminous data so we could examine specific (language) phenomena that occurred. (Published in NAACL: Demonstrations, 2018)

1.5 Structure of the Thesis

Having established the importance of cohesion (and hence cohesive devices) within discourse and SMT translation, and having highlighted the associated difficulties, we outline the structure of this thesis:

In Chapter 2 we explore a brief history of NLP looking at both its progression and various different MT paradigms, through to the modern day. In Chapter 3 we provide a review of cohesive devices. This gives a high-level view of various devices that cause problems for MT and shows why their omissions give rise to rough translations. Particular focus is also given to some important differences between Chinese and English that could potentially be a problem for MT. Chapter 4 presents a literature review looking at pertinent related work including an examination of some barriers in MT, and annotation of DMs. Chapter 5 is a corpus-based study of the divergence between the usage of DMs in Chinese and English highlighting important structural differences in composite sentences extracted from a number of parallel corpora. Some examples of how these cases are dealt with by popular SMT systems are also shown. Chapter 6 develops this and looks at ways to improve the translation of discourse connectives and discourse relations. This chapter also describes the effect of implicit elements on MT and looks at methods to identify them and make them explicit. It then details how we built upon these findings to create a usable prediction model that can predict implicitation in sentences from a source language.

Being able to visualise word alignments for SMT and general MT system output was of significant importance for this research. As such, we developed two comprehensive software systems that were used in much of our analysis. Chapter 7 shows our word alignment visualisation tool (WA-Continuum), and also details our Vis-Eval tool. WA-Continuum makes it easy to understand how words for each sentence have been aligned during the word alignment process. Vis-Eval enables the user to view the translation output of MT systems and compare multiple evaluation metric scores at both the sentence and dataset level, in fine-grained detail.

Chapter 8 provides the final conclusions and summary of our work, and includes some suggestions for future development.

Chapter 2

2 History and Overview of Machine Translation

Here we present a brief history of MT and NLP up to the modern day and we include how Chinese became a major language that was targeted for translation, initially due to defence and security projects.

2.1 A Brief History of Machine Translation

Using computers for automated translation is not a new idea. In 1949, the same year the People's Republic of China was founded, a man named Warren Weaver put forward a memorandum suggesting the possibility of using machines for translation (Mitkov, 2004), an idea perhaps stemming from the Enigma and Bletchley Park code-breaking successes during the war. Weaver, in his memorandum on translation (July 15th, 1949), suggested, that one language was just code form of another language using different strange symbols (Weaver, 1949; Koehn, 2013).

In 1950, the beginnings of NLP came about through an article titled Computing Machinery and Intelligence authored by Alan Turing and published in Mind, a philosophy journal (Rapaport, 2005). In the article, Turing discussed a test (now known as the Turing Test) which highlights the criteria to which a computer must adhere in order to produce artificial, intelligent (natural) language or behaviour, which is ultimately indistinguishable from that of a human.

Not long afterwards in 1952 the inaugural MT conference took place at the Mas-

sachusetts Institute of Technology (MIT) followed by the release of the first related journal, Mechanical Translation, in 1954 (Mitkov, 2004).

Significantly, 1954, was also the year when the first demonstration aimed at showing the viability of MT was presented to the public. The demonstration was presented jointly by IBM and Georgetown University (Hutchins, 2005) and is known as the Georgetown Experiment. During the demonstration over 60 sentences with 250 words and just 6 rules of grammar were translated from Russian to English. This was quite a limited experiment, but at the time it was significant, especially with the US interest in the actions of the (then) Soviet Union. Although this was a relatively small-scale experiment, its success attracted funding (Hutchins, 2005) and the enthusiasm fuelled the now infamous quote from the demonstrators who claimed that "within three or five years, machine translation would be a solved problem". More importantly, in 1955, the new interest in MT paved the way for the launch of the Machine Translation Research Project of Georgetown University. Conversely though, certain people were disappointed by the small scale of the experiment and even suggested that barriers such as semantic disambiguation were impossible to overcome using automated methods alone (Koehn, 2013).

The 1960s saw some success with computer systems such as ELIZA, an early example of simple NLP, where input was analysed for key words, generating responses based on these words and assorted reassembly rules (Weizenbaum, 1966). The software was known to give surprisingly realistic responses. Unfortunately, the enthusiasm and activities surrounding NLP and MT were brought to a standstill in 1966 with the release of the Automatic Language Processing Advisory Committee (ALPAC) report. The report was based on a study of the viability of MT, which showed that it was not cheaper or quicker than human translation because numerous certain tasks (e.g. post-editing) still had to be undertaken to produce a useful output (Koehn, 2013). Consequently the funding for MT, in the US at least, was significantly reduced and many projects halted.

Despite this, internationally there was some work that continued, which produced a small revival. For instance, in the late 60s there was a new lead in NLP activities. The book Computational Analysis of Present Day American English (Kucera and Winthrop-

Nelson, 1967) was published and it is essentially a deep analysis of the first well known (million-word) Brown corpus compiled at Brown University in 1963-64 (Francis and Cucera, 1964; Jurafsky and Martin, 2009). Some early commercial rule-based machine translation (RBMT) systems were also produced including Systran and Logos, the former being used in the US Air Force and in 1976 (the end of the Mao era) it was installed at the European Commission (Mitkov, 2004). Almost at the same time another system came into the fore and has since become one of the more successful early projects. It is known as the TAUM-MÉTÉO (or just MÉTÉO) system, developed by the University of Montreal and it was used (until 2001) to translate weather forecasts from English into French. Part of the success was largely down to the use of a specific domain (weather forecasts) and a sublanguage, which is easier to parse than a full language (Mitkov, 2004).

Up to this point the work appeared to focus on limited language pairs such as Russian and English or French and English. However, as technology advanced and the world became a smaller place (in terms of travel and communication) through globalisation and trade, translation of other languages (e.g. Japanese) saw an increased demand. This was especially true as personal computers started to become available. With an increase in the use of microcomputers or desktops came the development of simple computer assisted translation systems such as Trados (now SDL Trados) (Koehn, 2013).

In the 80s, as computational power increased further (Moore's Law) MT once again started to look viable. Example-based machine translation (EBMT) became popular, especially in Japan and many EBMT systems were built. Interlingua MT was also proposed, in which the source language segment was converted to a common in-between language and certain rules were then used to translate from the in-between language into the target language. The idea was that you only need one set of extra rules per additional language being added and that you would always be translating into and from a common middle language.

In 1986 the first issue of Computers and Translation (renamed Machine Translation in 1988) was published and the International Journal of Machine Translation soon followed in 1991 (Mitkov, 2004). In the late 80s IBM Research started looking at SMT, instead of

its word based translation, on the back of their successes with using statistical methods in speech recognition (Koehn, 2013). In the early 90s IBM also developed the Candide SMT system and the IBM statistical alignment models (1 to 5).

From 1989 – 1996 the Penn Treebank produced over 7 million words of part-of-speech (POS) tagged text. The materials used for annotation included IBM user manuals and Wall Street Journal articles (Taylor et al., 2003). At the time it was revolutionary and it became a leading example of POS tagging. The Penn Treebank is no longer in operation or being maintained, but because of the large amount of data produced it is still being used in NLP projects. Interestingly the tools and methods used were adopted to build the English Penn-Helsinki corpus (Kroch and Taylor, 2000).

The Defense Advanced Research Projects Agency (DARPA) also started making strides in speech and text processing working on automatic transcriptions of speech, initially with a 1000-word vocabulary, which soon expanded to 20,000 words and more (Olive et al., 2011).

In the late 90s DARPA started worked on transcribing radio news (TRVS program), incorporating some language analysis. The focus was mainly on English, but work on Arabic and Chinese was also subsequently included. Interest in automated Arabic translation was further solidified in September 2001 as a result of the 9/11 attacks on New York (Koehn, 2013).

Post 2004 saw an extension of DARPA with the introduction of the Global Autonomous Language Exploitation (GALE) program. Again this was essentially set up to cover extra requirements of the Defense Department. The idea was to create a system that not only could translate source material from many languages into English, but also could actually sort the information from the material, and separate relevant sources from irrelevant ones (Olive et al., 2011).

The funding poured into DARPA and GALE became a large part of the boost in the use of statistical methods in MT. In addition the ever-increasing availability of computer power and storage coupled with more and more digital resources (e.g. parallel corpora) has meant that developers can create more impressive, advanced and complex SMT sys-

tems.

Many MT systems are being developed across the globe and commercial systems include IBM, Microsoft and Google. The latter two provide free to use MT translation tools on the internet whilst IBM have developed Watson, an intelligent robot that responds to natural language. Watson was shown off to TV audiences in 2011 when it won first prize (1,000,000 dollars) on the Jeopardy game show. It processed questions given in natural language (having to deal with puns, synonyms, homonyms etc) providing answers from its extensive databanks without being connected to the internet. Watson has also been used in a call centre and is now being used to help fight lung cancer.

Systran has also moved into using SMT by incorporating it into its system alongside RBMT. This makes it a hybrid system that aims to take advantage of multiple models combining their relative strengths to improve translation quality (Systran, 2014).

Now SMT, despite being the dominant paradigm for a number of years, has taken a side step and has effectively been replaced by NMT, which is proving very successful.

2.2 An Overview of Machine Translation Paradigms

This section highlights a number of the more popular MT paradigms that have either been used in recent history or are indeed still being used.

2.2.1 Rule-Based Machine Translation

Even during the Georgetown Experiment in 1954 a number of grammar rules were used to assist with the translations. RBMT systems tend to use explicit rules provided by linguistic experts and large dictionaries that are coded into the program and used to deduce the required translation (Clark et al., 2013). In the early days Systran used RBMT technology in their system, which included a large set of rules alongside syntactic, semantic, and morphological material (Systran, 2014). There are many commercial MT systems that still use the RBMT concept (Clark et al., 2013), which may come down to the fact that a lot of the sophisticated linguistic rules and large dictionaries that have been amassed over the years are just too valuable to waste.

One of the main problems with RBMT is exceptions to the rule. Those exceptions can in turn have their own exceptions and so on, which can ultimately lead to contradictory rule creation leading to a set of very complex explicit rules.

2.2.2 Example-Based Machine Translation

The EBMT paradigm first appeared in the 80s (an early data-driven model), but gained real momentum in the 90s (Mitkov, 2004). EBMT systems use existing proven translations as a basis for the new translation. The input is essentially matched to its closest counterpart in a large database of existing example translations. The necessary (attempted) adjustments are then made for the words that differ from the input and its closest database match (Koehn, 2013), in order to try to produce a useful and reliable output. EBMT systems can in fact continue working with smaller and smaller translation chunks, incorporating the results in to the final output.

2.2.3 Transfer and Interlingual Models of Machine Translation

The Transfer model and Interlingual model share a common theme where alterations to some input are made before translation. The transfer model adjusts the input structure to mimic the structure of the target language (based on a set of rules) whilst the interlingual model converts the input into a middle pseudo-language, which is then converted into the target language. Whereas the former model requires a specific set of transfer rules for every language pair, the latter model aims to use simple syntactic and semantic rules to convert the extracted meaning into the target language (Jurafsky and Martin, 2009). The interlingual model is seen as more efficient and some systems, such as the KANT MT system, were built on this technology (Nyberg et al., 1997).

2.2.4 Statistical Machine Translation

SMT in its various forms is a widely studied MT model. Basic SMT does not make use of the normal linguistic data (Mitkov, 2004), but instead learns significant patterns that should not ordinarily occur by chance. The patterns and therefore the statistical models

based on them are developed by scrutinising a large data set of good translations referred to as parallel texts or bilingual corpora.

Commonly, individual words (and word groups) from the corpora are initially aligned, with phrases being extracted using some heuristics. Next, the probabilities that each of the items in one language map on to or correspond with the translations in the other language are calculated (Mitkov, 2004). Using this method the translation becomes a mathematical machine learning problem that does not rely on linguistic knowledge (also known as non-linguistic or anti-linguistic).

In around 20 years SMT became the dominating state of the art paradigm (Lopez, 2008) (although now it has, for the most part, been superseded by NMT) and has seen many changes, starting with word-based translation (WBT) (e.g. early IBM models). Through development phrase-based translation (PBT) became more widely used followed by syntax-based translation (SBT), hierarchical phrase-based translation (HPBT) and tree-based translation (TBT).

2.2.5 Word-Based Translation

IN WBT the usual atomic unit of translation is individual words. The early IBM models work on statistics collected from translated texts in a parallel corpus. IBM model 1 was very limited and the reordering process was poor. With each iteration of the IBM models came more power and complexity, but in general the translation process was broken up into small steps in a procedure called generative modelling, using word translations found under the lexical translation probability distribution (Koehn, 2013). The mappings worked well for one to one translations eventually allowing for some reordering. However, translations were based on what are now seen as relatively simple word alignment models (Clark et al., 2013).

Fast-Align (Dyer et al., 2013) (as discussed in Section 4.4) essentially uses a fast version of IBM Model 2, whereas Giza++ (Och and Ney, 2003) extends to using IBM Models 4 and 5. As such it is deemed pertinent to discuss the various IBM Models here.

Essentially, the IBM Models are a sequence of SMT models that increase in complex-

ity with each iteration. Early models started by using simple lexical translation probabilities, with later models incorporating reordering and word duplication processes.

The initial work at IBM looked at five models for SMT (Brown et al., 1993b):

- IBM Model 1, this consisted mainly of lexical translation probabilities that looked at the probability distributions of single words in a text. It worked to a degree, but had many flaws, including weaknesses around reordering, treating all reordering options as equally likely (Jurafsky and Martin, 2009).

- IBM Model 2, was similar to Model 1, but added the absolute alignment model (Clark et al., 2013), which was based on the actual positions of words in the input source and target sentences. IBM Model 2 simply consisted of the lexical translation step and the alignment step. The Fast-Align tool works at this level.

- IBM Model 3, added a fertility function to the process. This essentially means that single words in one language could be matched to multiple words in another and vice versa. This was considered to be a strong model for SMT in a word based paradigm (Jurafsky and Martin, 2009).

- IBM Model 4, introduced relative distortion (Koehn, 2013), which essentially allows for longer range reordering and makes use of word classes. It could handle situations where some words moved forwards or backwards whilst others stayed in the same place.

- IBM Model 5, the last of the original models, enhances Model 4, by further improving the alignment process and addressing deficiency (Brown et al., 1993b). Deficiency occurs when the model puts an output word into a position that has been filled. In theory multiple output words may be placed in the same position. Model 5 tracks the free positions and only allows words to be placed in those slots. Therefore the distortion model is similar to that in Model 4, but is focussed on free positions.

An IBM Model 6 was also created (also known as Model 4B), which was essentially IBM Model 4 combined with a Hidden Markov Model (HMM) (Mitkov, 2004), a first

order dependency. Whilst IBM Model 4 aims to predict the distance between target language token positions, the HMM looks at the distance between source language token positions (Koehn, 2013). Combining both approaches (in a log-linear manner) meant improved results over the original 3 models.

All said and done, the main weaknesses with word-based translation are that words are not the best atomic unit for translation as there can be one to many mappings across language pairs. In addition, when focussing on single words all contextual information is lost, so no context can be utilised in the translation process (Clark et al., 2013).

2.2.6 Phrase-Based Translation

PBT was considered by many to be the leading model in SMT for a long time (Marcu and Wong, 2002) and, to a degree, related to EBMT (Koehn, 2013). With its introduction, systems were able to learn both phrasal and lexical alignments. This created a method far more powerful than learning lexical alignments alone, and consequently translation quality improved. One of the shortcomings of WBT was when there were many to one (or vice versa) mappings of words across a language pair then the translation broke down and all context was lost. Using phrases as the smallest atomic contiguous unit (n-grams) removes some translational ambiguity, and helps with some reordering problems (Clark et al., 2013). The most basic PBT model is the noisy channel approach (Shannon, 1948; Brown et al., 1993a) as shown in equation 1.

For equation 1:

e is the best English sentence

F is the foreign language sentence

P is probability.

The best English sentence is the one that has the highest probability, that is: $P(E|F)$

Rewriting this using Bayes' rule gives:

$$e = P(E|F) = \frac{P(F|E)P(E)}{P(F)} \quad (1)$$

The P(F) in the denominator can be disregarded (Specia, 2013), as we want the best English sentence corresponding with the given foreign sentence. This means F is a fixed constant from the source text (e is essentially independent of $P(F)$). This gives:

$$P(F|E)P(E) \quad (2)$$

Equation 2 is made up of two main components:

1. a translation model $P(F|E)$
2. a language model $P(E)$

The translation model appears to be the reverse of $P(E|F)$. This is due to 'F' being taken as some noisy code version of an English sentence or phrase (as per Weaver, 1949). The task then becomes one of finding the hidden 'E' sentence that produced the polluted 'F' sentence (Jurafsky and Martin, 2009). This last step is 'solved' through using a decoder that aims to produce the most likely 'E' sentence, when given 'F'. In the case of the PBT model, the probabilities are usually calculated using phrases rather than individual words.

2.2.7 Syntax-Based Translation

One shortcoming of PBT is that it can often fail to capture long range movement. SBT aims to overcome this issue by modelling a deeper level of structure (Chiang, 2007). A syntax is incorporated into the SMT system, which should produce better output. One of the main barriers to this approach though is translation speed. Although, when using SBT, improvements in translation have been observed the speed of the translation is appreciably slower. In addition, syntactic annotation is not normally marked up in sentences and adding it (automatically) would make models more complex and difficult to manage. As such, SBT can be messy when using parallel corpora leading to complexity in the model that is difficult to address (Koehn, 2013).

2.2.8 Hierarchical Phrase-Based Translation

This is essentially an SMT model that examines phrases within phrases (i.e. sub-phrases). It also trains on parallel corpora, and combines ideas from both PBT and SBT (Chiang, 2007). It has been observed that moving to hierarchical structures can markedly improve the quality of translation (Chiang, 2007) within certain frameworks. Additionally, numerous researchers have found that incorporating SBT into their models shows improvements and HPBT has become a popular alternative to the flat PBT paradigm (Clark et al., 2013).

The key advantage to HPBT is that it can tap into the true power of PBT; that is, phrases can be used to learn word reordering. By extension, the same methodology used to learn the reordering of words can be applied to learn the reordering of small phrases (or sub-phrases). This process can be applied even when simpler distortion models are used. The problem is not with the use of a plain distortion model, but rather one of identifying the basic units for translation.

The atomic unit in HPMT is a hierarchical phrase, which primarily is made up of words and sub-phrases. This means it can capture translations that are beyond the standard short phrase (usually tri-grams) used in traditional PBT. This can then remove some of the limitations commonly associated with PBT and and reduces the problem down to 'simpler' grammatical rules.

(10) 那个戴帽子的男人叫大卫。(Chinese)

that wear hat [de] man name David

(Literal)

The man wearing the hat is called David.

(typical translation)

In Example 10 we can see '叫大卫。' (called David) at the end of the Chinese sentence can remain in place and so no reordering is required. The same goes for '那个' (the/that) at the beginning of the sentence.

In fact the main difference is the reversal '戴帽子' (wearing the hat) and '男人' (man) around the character '的'. Here '的' is used as a linking word that links descriptive words,

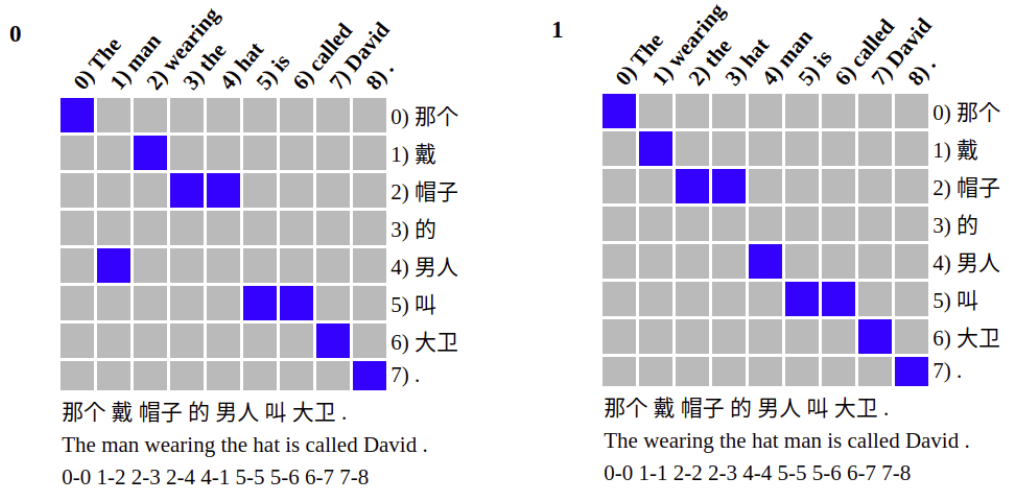

Figure 1: Visualisation of word alignments showing the literal order of the English aligned to the Chinese vs the natural order.

phrases, or clauses to the noun they describe, in this case ‘the man’.

For this sentence pair the reordering rule is simply a swap or reversal of the descriptive phrase and the noun. That is, in the Chinese the descriptive phrase is to the left of the noun, whereas in the English it is to the right of the noun.

In Figure 1 the grid numbered 0 shows a typical correct alignment of words according to the natural order of the given sentences in both languages. In grid 1, the English has been changed to follow the more literal order of the Chinese.

A true diagonal line (as per Figure 5 in Chapter 7) would show a sentence pair that needs little to no reordering. Here the blocks in grid 1 are almost a true diagonal from top left to bottom right, and if the English could remain as a literal translation then no reordering of phrases is required.

However, the true alignment we need is shown in Grid 0. As can be seen the blue blocks in rows 0, 5, 6, and 7 are the same in both grids, meaning that no reordering is required for those rows. The action is all in the middle of the grids. The block in row 4, column 1 (grid 0) has to be moved back from column 4 (in grid 1), and the blocks in rows 1 and 2 move across 1 place as a result. This in effect is a visual representation of the reordering taking place. The descriptive phrase in rows 1, and 2, moves to the other side

of the noun in row 4.

This is not a case of one word being swapped for another, but more a case of a word (noun) being swapped with a (descriptive) phrase.

2.2.9 Tree-Based Translation

Tree-based SMT includes a variety of different tree models that aim to overcome the weaknesses of earlier SMT models such as coping with long distance dependencies required with numerous language pairs (Specia, 2013) and having too much elasticity with respect to word reordering (Wu, 2005). The tree models also enable the inclusion of syntactic and semantic information (Clark et al., 2013). There are numerous tree models including: HPMT (as discussed), tree to string, string to tree and tree to tree (Nguyen et al., 2008). Tree models include the use of syntactic trees, which focus on syntactic relationships between words and phrases (Koehn, 2013). Representing syntax with trees is called grammar, and there are numerous different grammars such as: tree-adjoining grammars (TAGs), tree-insertion grammars (TIGs) and synchronous TIGs (STIGs) to name but a few (Clark et al., 2013). Synchronous grammars essentially create both a source and target language tree at the same time and a synchronous tree-substitution grammar can create pairs of non-isomorphic trees that help overcome the reordering problems of flat models (Koehn, 2013).

A detailed analysis regarding the entire range of trees and all possible grammars is outside the scope of this work. The discussion here is intended to highlight some of the positive aspects that trees have over flat systems. Clearly there are other factors to consider that add complexity to implementing tree models, such as decoding, aligning sub-trees, pruning and adding syntactic annotation. It is worth noting however that 'basic' HBMT systems that combine phrase-based models and trees (as discussed earlier) have shown promise in terms of improved translation quality (Chiang, 2007) and are indeed worth considering for use as a base system.

2.2.10 Neural Machine Translation

At the time of writing NMT is an emerging paradigm, which has led to some major improvements over previous SMT and RBMT translation systems (Wu et al., 2016). Whilst SMT is still widely used and produces strong results it appears that NMT has overtaken it and become the gold standard, especially as previously unavailable requisite technology (e.g GPUs) has now become mainstream.

Whilst bilingual corpora are often still used for training the systems, other processes (such as initial word alignment) can be dropped entirely and the models still work. New methods, such as the inclusion of attention mechanisms have been introduced. These can be quite powerful and network architectures such as Recurrent Neural Networks (RNN) (and the enhanced Long Short Term Memory (LSTM) variants) have shown great promise. As with SMT, it is anticipated that continual improvements such as these will be made to each new iteration of NMT models and the cycle will continue bringing better and smoother translations with each step. A thorough discussion of NMT models and neural network architectures is outside the scope of this thesis.

Chapter 3

3 Background on Cohesive Devices

Chapter 1 briefly introduced cohesion and highlighted commonly identified cohesive devices as described in the observations of Haliday and Hasan (1976). Whilst the devices cannot always be separated exclusively into exacting distinct categories, they will be discussed in terms of grammatical and lexical cohesion in order to provide clarity. The devices in their respective categories are:

1. Grammatical Cohesion
    - Reference
    - Ellipsis
    - Substitution
    - Conjunction
2. Lexical Cohesion
    - Collocation
    - Reiteration

The remainder of this section is dedicated to presenting an overview of cohesion from a linguistic perspective.

3.1 Grammatical Cohesion

Grammatical cohesion is divided into four main parts: reference[4], substitution, ellipsis and conjunction. Reference covers a particularly wide range of words, which are based on culture and knowledge (homophoric referencing), being in a particular situation (exophoric referencing) or linking within text (endophoric referencing). Table 1 illustrates the given distinctions.

Table 1: Grammatical cohesion and cohesive devices.

Grammatical Cohesion	
Reference	Homophoric (knowledge based)
	Exophoric (situational)
	Endophoric (textual) Anaphoric (refers backwards) Cataphoric (refers forwards)
Ellipsis	
Substitution	
Conjunction	

3.1.1 Referring

Referring is the means used for relating one element of text to another enabling better interpretation of the information. The elements can be present within the text (endophoric) or be outside of the text (exophoric, homophoric). Homophoric and exophoric referencing are both briefly discussed here before detailing endophoric referencing, which is more pertinent to text/discourse.

3.1.2 Homophoric Reference

Homophoric reference is where the identity of the item being spoken or written about can be ascertained by widely accepted knowledge and cultural behaviour in general, rather than referring to the local context within the text.

- **the** moon was so bright...
- all over **the** world...

[4]Issues of reference, deixis and function are discussed from the perspectives of text and discourse analysis, rather than speech.

In the previous two given examples the word '**the**' is actually referring to the moon and the world in general. There is no need for either item to be introduced in the text for the reader to understand the meaning or context. The reader would not normally need to ask 'which world?' or 'which moon?'. Contrast that with the next sentence:

- **the** scene was utter devastation...

In this latter example the reader must be referred to the scene at some point in the text to understand the context, which in this case is likely to be some sort of news story covering an accident. Otherwise the reader may be left wondering 'what scene?'.

3.1.3 Exophoric Reference

Exophoric reference is perhaps slightly more localised than homophoric reference, but it still goes beyond the boundaries of the text. From the text alone the identity of the item being discussed is difficult to ascertain without actually being in the situation.

- ...please pick **it** up and place **it** over **there** next to **that** desk...

In the example sentence: 'it'[5], 'there' and 'that' are impossible to work out without actually being in the situation or seeing it.

3.1.4 Endophoric Reference

With regard to cohesion within the scope of a text, the main pertinent subtopic of referring words is endophoric reference. Endophoric links are used to tie the internals of the text enabling the reader to refer clearly and concisely to previous or impending discourse segments, whilst concurrently avoiding repetition. Such referencing is called deixis or deictic referencing. Deictic referencing, which refers back, is anaphoric deixis, and that which refers forwards is cataphoric deixis (Table 1).

[5]In the given example the second 'it' refers back to the first 'it'.

3.1.5 Anaphora

Anaphora resolution (referring) is the process of establishing the antecedent of an anaphor (often a pronoun) and is a known difficulty for MT (Jurafsky and Martin, 2009). Even when dealing with just pronouns (pronominal anaphora), their usage and distribution across different languages can vary (Russo, 2011), and this is before considering number and gender agreement. Past corpus studies have highlighted how this can be a problem for SMT systems leading to possible mistranslations of pronouns (Hardmeier, 2014).

The nature, function, usage, and distribution of generic pronouns varies between languages, and many issues, such as number agreement, which relies on information connected to the antecedent[6], need to be dealt with accordingly.

In English, personal pronouns are often used to refer back to: nouns, noun phrases, or other pronouns. Demonstrative pronouns (this, that, these, those) work in a similar way and can often refer back to the entire preceding sentence. Determiners (specific and general) are also often used to refer back (e.g. 'the', 'a' and even quantifiers). Auxiliary Verbs (e.g. be, do, have) work in a similar fashion to pronouns, but they refer to the verb phrase to avoid repetition:

- I saw a tiny spider and ran away screaming, but unfortunately I fell over and landed in a muddy puddle. **This** was embarrassing.
- I have five brothers, **three** live in London.
- I was going to help him fix his car. Unfortunately I was unable to **do** so because of the weather.

Other words including so, not, and such are also often used to refer back (e.g. if so then... / if not then...). Of these 'such' is a very interesting case of substitution, it functions in many different ways and can behave as a determiner, a predeterminer and even an adjective:

- China burns a lot of coal. Using **such** fuel causes pollution. (determiner)

[6] A word, phrase or clause that is replaced by a pronoun (or other substitute) later or earlier in text - usually the following sentence.

- The bridge collapsed. Fortunately **such** an event is rare. (predeterminer)
- He can often be very annoying. This was one **such** moment. (adjective)

3.1.6 Cataphora

Although cataphoric reference is not as common as anaphoric reference it still occurs with a degree of frequency. There is some overlap with the ideas presented in the anaphora section and so, for brevity, just a few examples are given here as an illustration.

- **following/next/below** - Please use the following address.
- **this/these** - You may not believe this, but it will be sunny tomorrow.
- **pronouns** - When he returned home, John noticed that the door was ajar.

3.1.7 Ellipsis

This is the omission of one or more words, which would be, in the case of text, supplied mentally by the reader. Ellipsis essentially encourages the reader to 'fill in the blanks' with the correct item, which can be ascertained from the surrounding information.

- John would love to visit China, but he cannot afford to (... *visit China*).

3.1.8 Substitution

Substitution is slightly different to ellipsis in that a word (or phrase) is replaced (substituted and not omitted) by a more general filler word such as do, not, one or so. This process avoids the need for repetition.

- Dan loves strawberry ice-cream. He has **one** every day. (Halliday and Hasan, 1976)

In the example 'one' is a substitute for ' strawberry ice-cream' meaning the longer item (strawberry ice-cream) is not repeated, but the understanding is maintained.

3.1.9 Conjunction

This final device for grammatical cohesion is very important as it provides a relation between discourse segments through the use of specific discourse markers.

In English there are many discourse markers that can be used and they express different relationships. Table 2 shows a number of such relationships combined with their respective markers and sentence position[7][8] (Chuang, 2017).

Table 2: A selection of different discourse markers and their relations.

Type of relationship	Sentence connectors	Position within the clause or sentence
Adding something	Moreover; In addition; Additionally; Further(more)	Initial
Making a contrast	However; Yet, On the other hand	Initial
Making an unexpected contrast	Although; Despite the fact; Even though; Regardless of	Initial or starts subordinate clause
Saying why	Because; Since; As; Insofar as	Initial or starts subordinate clause
Saying what the result of something is	Therefore; Consequently; As a result; Hence; Thus; Accordingly; For this reason	Initial
Expressing a condition	If; in the event; as long as; Provided that; Assuming that; Given that	Initial or starts subordinate clause
Strengthening your argument	On the contrary; As a matter of fact; In fact; Indeed	Initial

However, there are times when simple conjunctions can be used (e.g. and, but, or). In addition there are other types of patterns as well. One of particular note is correlative conjunctions such as: ...not only... but also; neither... nor...; both... and... .

One final pertinent part of conjunction is that of linking wider discourse segments, particularly paragraphs. There are many types of paragraph relation, but the three main ones (Chuang, 2017) are:

1. Reinforcement of ideas - A further example...

2. Contrast of ideas - This argument is not however...

[7]Sentence connectors can begin a clause or sentence after a semi colon.
[8]Some sentence connectors can be placed in different positions within a sentence.

3. Concession - Although the arguments are...

The illustrated links and ties to clauses, sentences and paragraphs, if used correctly, appear to be essential to the creation of unified and connected text. If this is indeed the case then by the same token modelling such relations should, in the very least, serve as an aid to guiding automated translations.

3.2 Lexical Cohesion

Grammatical cohesion covers a complex variety of frequently overlapping cohesive devices that cannot always be completely distinguished. Lexical cohesion is altogether an easier category to discuss as it is essentially the selection of vocabulary to create links for unified text; that is, cohesion is directed by relations between words across textual units (Jurafsky and Martin, 2009).

The two main parts of Lexical cohesion, as shown in Table 3, are reiteration and collocation.

Table 3: Lexical cohesion and cohesive devices.

Lexical Cohesion	
Reiteration	Repetition
	Synonymy
	Antonymy
	Hyponymy
	Meronymy
Collocation	

3.2.1 Reiteration

The broader concept of reiteration includes direct repetition of items, as well as a more subtle, disguised repetition though generalisations - using super-classes of words, or specifications using subclasses.

Synonymy is the idea that two or more words can portray the same meaning, with or without identity of reference (Halliday and Hasan, 1994). For example, noise and sound have the same level of generality and are synonyms in the narrower sense. On the

other hand birds referring back to blackbirds becomes a superordinate term (Halliday and Hasan, 1994). The remaining devices of reiteration (antonymy, hyponymy and meronymy all perform a similar function). Antonyms are words with opposite meanings; hyponyms are words that are a subclass of the superordinate more general word; meronyms are used to show a part-whole relationship, where the part is a constituent piece of the whole.

- Synonym - beautiful and attractive are synonyms
- Antonym - good is an antonym for bad
- Hyponym - pigeon is a hyponym for bird[9]
- Meronym - finger is a meronym[10] (constituent part) for hand

3.2.2 Collocation

Collocations are familiar collections of words (similar to a phrase) that usually appear together and tend to only convey meaning through association (Nordquist, 2014) and are often specific to a culture or context. Table 4 shows a miscellaneous range of well-known collocations[11], but there are many more.

Table 4: A selection of miscellaneous collocations.

Time	Business English	Classifying Words
bang on time	annual turnover	a ball of string
dead on time	bear in mind	a bar of chocolate
early 12th century	break off negotiations	a cube of sugar
free time	cease trading	a bottle of water
from dawn till dusk	chair a meeting	a bunch of carrots
great deal of time	close a deal	a pack of cards
late 20th century	close a meeting	a pad of paper
make time for	come to the point	
next few days	draw a conclusion	
take your time	make a loss	

[9] Bird is the hypernym for pigeon.
[10] Holonym is the opposite of meronym - that is, hand is the holonym for finger.
[11] https://www.englishclub.com/vocabulary/collocations-common.htm

3.3 Chinese vs English

Sections 3.1 and 3.2 gave a high-level view of different aspects of cohesion and showed how much information various devices hold and hence why they cannot simply be ignored. Here we look more closely at some differences between Chinese and English, which can be vast and thus cause problems for SMT. Some of the main differences are:

- word order and sentence structure
- verb forms
- time and tense
- relatives
- articles
- pronouns
- gender and number

3.3.1 Word Order and Sentence Structure

Chinese sentences are often topicalised, where a sentence starts with a detached subject or object (Swan and Smith, 2004) such as: 老人，必须尊重(literally) 'old people, must respect', which in English would be more akin to 'old people - we must respect them', better translated as: 'the elderly must be respected'. Already this shows that words can be routinely omitted in one language, but not the other.

In addition to omissions, word order is a known problem in MT for many language pairs. Chinese has vast word order differences from English. Topicalisation is a good example of this, as is the adjustment of word order as a substitute to inflection. There are other common word order differences from English, for example, in Chinese, statements and questions can have an identical word order, but often with an additional question particle added at the end.

3.3.2 Verb Forms, Time and Tense

Chinese is classed as a non-inflected language. So where in English the verb is changed, in Chinese the effect is achieved by using word order, adverbials, and context (Ross, 2011; Ross and Sheng Ma, 2006). With no inflection it can be hard to determine what tense is required for an isolated segment of a Chinese sentence. Chinese also does not express time or tense as in English. For example, in Chinese there is no verb conjugation to express time relations. Again this is all information that can be implicit causing additional problems for SMT.

3.3.3 Relatives

Relatives are often treated very differently in Chinese and English. For example, zero pronoun structures such as:

'the house we wanted was too expensive'

can be translated as:

我们想要的房子太贵了

'we want house too expensive' (literal)

3.3.4 Articles

There are no articles in Chinese (Swan and Smith, 2004), although some approximations are sometimes evident for 'a' (一个- one) , and the characters 这个(zhège - this) and 那个(nàge - that) often serve a similar function, but are not a replacement for English articles. The word 'the', possibly one of the most commonly used words in English, has no direct equivalent.

For example,

'Let's go to the cinema.'

我们去看电影吧。

'we go look film [suggestion particle].' (literal)

Again, a quick look at the given sentence shows potential problems for MT, where articles should be inserted into the English.

3.3.5 Pronouns

English uses pronouns much more than Chinese. Chinese is considered to be a pro-drop language (Swan and Smith, 2004), where pronouns are dropped when they may be understood, especially from context. This is discussed in detail in Chapters 5 and 6. Ultimately though, similar to articles, it can cause situations where information is implicit in one language, but required in the other. This is especially an issue with isolated sentences and segments.

3.3.6 Gender and Number

There is no real clear distinction in gender in spoken Chinese as the characters for he, she, and it (他， 她， 它), whilst all different, are pronounced in the same way. This is not normally an issue in text, but it is possible that corpora from speech (e.g. TED or TV) may have gender errors.

Number is also problematic as there is not normally a way of expressing plurals in Chinese[12]. Some sort of quantifier is therefore needed to be sure of any plurality (Tung and Pollard, 1994; Po-Ching and Rimmington, 2010). Clearly this can be problematic if the quantifier is not in the current segment.

The described elements are often intertwined and cannot just be routinely dropped during the translation process. Information from these elements is often key into gaining a wider understanding from text. This overview is used to highlight the vast divergence between Chinese and English. Of the discussed elements, for this thesis, we pay particular attention to pronouns and discourse connectives.

[12]们as in 他们- them - can be used in some cases. Also as in 我们去看电影吧.

Chapter 4

4 Literature Review

SMT was considered to be the dominant form of MT for well over a decade (Specia, 2013). Word-based models were developed in the 1980s (IBM Candide) and then phrase-based models took over as the gold standard. Despite continual improvements in the translation quality output from SMT systems they are still considered as being unable to deliver human quality translations especially with complex sentences and divergent language pairs. Here we outline some of the barriers to MT.

4.1 General Barriers to Machine Translation

Much research and funding has gone into developing MT systems, but yet there are still many barriers that stand in the way of fully automated high quality MT. Some of the major difficulties are discussed here including:

- The availability of substantial, quality parallel texts on which to train the models (becomes less of an issue with greater storage capacity)
- Morphology across languages
- Word reordering
- Named Entity Recognition and unknown words

Each highlighted item by itself can present challenges for MT systems, but when translating across divergent languages many of the issues can arise simultaneously.

4.1.1 Resource Availability

Translation of low-resource language pairs is still considered an open problem in SMT (Lopez and Post, 2013), but with the increasing availability of data and computing power the size of this problem is constantly being reduced. Chinese, for example, was once considered a resource poor language in terms of available corpora, but more recently a lot of work has gone into developing such corpora and now there are numerous useful data sets. For example the ones detailed here are used for much of our analysis:

- Basic Travel Expression Corpus (BTEC): This corpus is primarily made up of short simple phrases and utterances that occur in travel conversations. For this study, 44016 sentences in each language were processed with over 250000 Chinese characters and over 300000 English words (Takezawa et al., 2002).

- Foreign Broadcast Information Service (FBIS) corpus: This corpus uses a variety of news stories and radio podcasts in Chinese. For this study 302996 parallel sentences were used containing 215 million Chinese characters and over 237 million English words.

- Ted Talks corpus (TED): This corpus is made up of approved translations of the live Ted Talks presentations[13]. This corpus contains over 300,000 Chinese characters and over 2 million English words (Cettolo et al., 2012) spread across 156805 parallel sentences.

- Multi-UN corpus (UN): This is a parallel corpus (for 6 languages) using data extracted from the United Nations Website. It includes over 220 million words in English and over 629 million Chinese characters in 8.8 million parallel sentences

4.1.2 Morphologically Rich Languages

Simply put, morphology in linguistics revolves around the way words are formed and their relationships to other words that exist within the same language. In English, mor-

[13] http://www.ted.com, and WIT3: https://wit3.fbk.eu/, both accessed July 2019.

phology looks at the roots of words as well as affixes, bases (word stem), inflections, and morphemes (the smallest unit of language).

The English word '**kind**', for example, can stand alone as a simple stem word, or through the use of affixes can become a complex word with a different meaning, such as '**un-kind-ness**'. The number and nature of such possible changes are what give a language its morphological flexibility (or richness).

Needless to say, many languages (and by extension language pairs) do not contain the same morphological flexibility (also see Chapter 3). A morphologically rich language (MRL) may have a free word-order or use various forms in their words (inflection). A MRL may also contain a greater range of domains or registers than a non MRL. For example German has a greater variety of registers than English and consequently has a freer word order. Chinese, despite generally having a SVO word order, also has a more flexible word order than English. English on the other hand uses more inflection than Chinese, but is still classed as 'weakly inflected'. English is not considered to be a MRL. Ultimately the differences between the morphological states of various languages present a challenge to SMT.

4.1.3 Word Reordering

Word reordering is one of the hardest problems in MT (Koehn, 2013). The issues created by differences in word order between languages vary depending on the language pair. For example, movement, between English and French (adjective-noun reordering) is, in many cases, good enough (although not solved). For German longer range movement, especially of verbs, is often required, and can be difficult (Clark et al., 2013). Shorter movement is also often good enough for Chinese and English (Koehn, 2013), but the flexibility of word order in Chinese poses an extra level of difficulty for reordering models.

4.1.4 Word Sense Disambiguation

Word sense disambiguation in SMT is also a heavily researched topic. WSD primarily revolves around identifying the sense of ambiguous words based on their context (Specia

et al., 2005). A simple example of this is the word **drive**. Without any surrounding context **drive** could take on a number of meanings. The seven examples given here demonstrate how PBT should outperform WBT by including words that are close by and at the very least provide some context.

- A nice drive through the countryside (a journey)
- I've had a new drive put in (a driveway to a house or a computer hard drive)
- I drive to and from work every day (operating a vehicle)
- Drive (the name of a film released in 1998 and again in 2011)
- This mess will drive me crazy (leading to a state of annoyance)
- Wow, he can drive that golf ball over 500 yards (to hit, throw or propel something)
- She has a lot of drive when it comes to studying (an urge or attitude)

4.1.5 Named Entity Recognition

Named entities (NEs) are an essential part of sentences in terms of human understanding and readability and recognising them is known as Named Entity Recognition (NER). To that end it is important that high quality MT systems correctly identify and translate NEs that occur in the input. The mistranslation, or indeed dropping of NEs can significantly impact translation output, although evaluation scores (e.g. BLEU) may not be adversely affected (Singla and Agrawal, 2009; Hermjakob et al., 2008). SMT systems were considered to be bad at translating names (Hermjakob et al., 2008) especially rare names. Name translation is a hard problem, especially when considering the wide variety of names across all languages and all domains. Singla and Agrawal (2009) illustrate this by example of a Chinese translation:

> Chinese: 27日中午, 他们已被安全转移到普吉岛。
>
> English: 27 noon , they have been shifted to safe places to 3pm .

The last three characters in the Chinese sentence 普吉岛(pǔ jí dǎo) can be translated as 'Phuket' or more specifically 'Phuket Island' (岛dǎo = island). However, the English translation drops the name altogether and just refers to the safe place.

Inputting the same sentence into Google Translate (2014) produces the output:

'27 noon, they have been moved to safety in Phuket.'

Clearly, **Phuket** has become a recognised NE, and in the case of Google Translate it is dealt with correctly[14]. This could suggest that the NER modelling technology has improved and through respective research (Singla and Agrawal, 2009; Zhang et al., 2013) plus the increased availability of data (especially for Google), NER has become a useful component technology that can be incorporated into SMT systems. NER is now able to better assist with translations of NEs of person, location, and organisation (Finkel et al., 2005), improving the overall translation quality.

4.2 Barriers to Translating Chinese and Using Discourse Relations

Chinese and English are considered to be a very difficult language pair (Koehn, 2013). Chinese writing is classed as logo-graphic or ideographic with sentences being made up by a number of non-segmented ideograms, in this case Chinese characters, which carry a certain meaning. This provides an extra barrier to translating Chinese texts. Here we examine segmentation issues and other problems with translating Chinese and working with discourse connectives.

4.2.1 Word Segmentation

In NLP (more precisely text processing) Chinese word segmentation is a very important task as there are no word delimiters (spaces) between words (Li and Yuan, 1998) although common punctuation is used. The accuracy of such segmentations is paramount as translations rely on the initial quality of segmentation to split the sentence into the correct words. Without accurate segmentation additional ambiguity is introduced and can cause problems (Olive et al., 2011).

[14]It could be argued that with so much access to data, Google's models have **Phuket** as part of their standard vocabulary.

Much research has gone into word segmentation broadly falling into two categories: lexical knowledge based (e.g. maximum length matching) and linguistic knowledge based (Li and Yuan, 1998). The GALE project (as part of DARPA GALE) sought to address such tokenisation and ambiguity issues with MT of text by making it a primary focus of their research (Olive et al., 2011). To do this the project concentrates on two main inputs – speech and electronic text (with MT from text being the centre of the project) looking specifically at transcription, translation and distillation of two languages (Chinese/Arabic) (Olive et al., 2011). A range of data types are used including news broadcasts, newsgroups blogs and of course resources from the Linguistic Data Consortium (LDC)[15]. The LDC, formed in 1992, has a rich and unparalleled collection of foreign text and media. Initially it was tasked with the role of acting as a repository of data, but now is responsible for managing the distribution of numerous approved corpora and various language resources. The GALE program became the largest resource creation project for the LDC and the research findings will be a significant source for future work at the LDC (Olive et al., 2011).

Over a five year period the GALE project produced a vast array of useful word alignment corpora. For example, the Chinese corpora incorporated character-based files for tokenisation (produced using automatic word segmentation tools). A manual error correction step was applied to the process in order to improve results.

Ultimately the best chosen model was a character-based model where each character (rather than a word) is classed as a separate token. The most success in GALE MT using the said word alignment has been achieved using statistical corpus-based approaches, which allow the addition of linguistic constraints (features). However, it has been observed that even though word segmentation is a prerequisite for automatically translating Chinese, error free segmentation of Chinese text is not yet possible (Xu and Bock, 2011) and, as such, segmentation ambiguity or inconsistency is still a big problem. That said, approaches for overcoming this barrier have been studied and significant performance gains have been recorded (measured by BLEU and TER) when using character-based

[15] https://www.ldc.upenn.edu/

translation trained with segmentation rather than word-based translation using the same tools (Xu and Bock, 2011).

4.2.2 Annotating Discourse Markers

With respect to DM usage, much work has also contributed to the field. Contributions include: annotating DMs in a treebank and in multilingual corpora; translating connectives in general; identifying and classifying connectives; and, to some extent, observing the implicitation of discourse connectives in MT.

The COMTIS project at IDIAP is closely connected to the DARPA GALE project (IDIAP was a contributing partner) and it has a strong focus on both Arabic and Chinese. A study on translating English discourse connectives (DCs) into Arabic (Hajlaoui and Popescu-Belis, 2013) showed that some DCs in English can be ambiguous signalling a possible variety of discourse relations. Clearly, if such ambiguity is not captured correctly when the discourse relations are translated, the likelihood of incorrect translations being produced is increased. However, other studies have shown that sense labels can be included in corpora and that MT systems can take advantage of such labels to learn better translations (Pitler and Nenkova, 2009; Meyer and Popescu-Belis, 2012).

The Penn Treebank, although no longer being developed, has been extended through the Penn Discourse Treebank project (PDTB) which adds annotation to English discourse connectives. Both the Treebank and the PDTB project were supported by DARPA. The Chinese Discourse Treebank (CDTB) is a project that adds an extra layer to the annotation in the PDTB (Xue, 2005). The CDTB project focuses on DCs that connect discourse relations in either a structural or anaphoric way. Structural relations use subordinate and coordinate markers that link close/neighbouring segments, such as clauses. Anaphoric relations usually have one discourse adverbial linked with a local argument, leaving the other argument or information to be established from elsewhere in the discourse. Pronouns for example, are often used to link back to some discourse entity that has already been introduced. This essentially suggests that arguments identified in anaphoric relations can cover a long distance and Xue (2005) suggests that one of the biggest challenges for

discourse annotation is establishing the distance of the text span and how to decide on what discourse unit should be either included or excluded from the argument.

There are also some additional challenges such as discourse sense disambiguation; the style and type of a discourse relation, for instance - is it implicit or explicit?; and possible DM variants. Table 5 (Xue, 2005) shows discourse connectives that can be used interchangeably without significantly altering the meaning. The numbers in each row indicate that, generally[16], any marker from (1) can be paired with any marker from (2) to form a compound sentence with the same meaning.

Table 5: Examples of discourse connective variation.

English	Chinese Discourse Connectives
although / but	(1) 虽然，虽说，虽 (2)但，还是，可是，却，然而，不过
because / therefore	(1) 因为，因，由于 (2)所以
if / then	(1) 如果，假如，若 (2)就
therefore	因此，于是

Studies in categorising discourse relations have been undertaken, and found that DMs in Chinese are part of an important language feature required for computational analysis. Unfortunately, recognition of such markers is more difficult in Chinese than in English (e.g. fewer word boundaries) and so analysis of Chinese syntax is generally more complex (Tsou et al., 1999).

DMs are essential for discourse analysis as they help signify the sequence of discourse segments and their respective relations (Table 6), which are used by the authors as structural and cohesive clues. It has been argued that results from discourse analysis can be applied to solve problems in NLP and thus have an important role (Forbes et al., 2001). For example, in some cases (e.g. Chinese newspapers) DMs appear in over 60% of sentences (Tsou et al., 1999), which highlights their significance.

As DMs form a significant part of discourse, accurately annotating them (and re-

[16] As always there may be exceptions, depending on the exact context at the time. For example formal vs informal register, but the general case holds true.

Table 6: Examples of discourse markers and their relations (Tsou et al., 1999)

Relation	DM Example	English Translation
Adversativity	虽然。。。但是	though...but
Causality	因为。。。所以because...	therefore/so
Concession	即使。。。	even...yet
Conjunction	以便。。。一边	on one hand...on the other hand
Deduction	既然。。。那么	if so...then
Degree	越。。。越	the more...the more
Disjunction	或者。。。或者	either...or
Intentionality	以使	so that
Necessity	只要。。。（才）	only if
Progression	不但。。。而且	not only...but also
Sufficiency	如果。。。那么	if...then

moving ambiguity) is essential for improving automated translation tools, and many approaches have been proposed including Naïve Matching and decision tree classifiers giving an accuracy rate of around 80% (Tsou et al., 1999), suggesting further linguistic analysis is required.

4.2.3 Lexical and Grammatical Cohesion

Here we examine some important and more recent developments that are closely connected with this research. A lot has been achieved by the DARPA GALE project and their partners (e.g. IDIAP), but there are other studies, such as the COMTIS project (again IDIAP) that have a strong focus on DMs and discourse relations using English and French as the main language pair.

The importance of DMs and their respective relations has already been established and some examples of attempts to annotate them have been shown. For the sake of cohesion the DMs have to be accurately interpreted, but this is a difficult task. Current SMT systems often struggle with correctly identifying lexical items such as pronouns and DMs. As a result the COMTIS project aims to expand the current SMT models through examining inter-sentential relations (Popescu-Belis et al., 2011, 2012).

Annotating discourse connectives (e.g. because) is a difficult task. While many lan-

guages contain some set of discourse connectives there is a great variation in the number available to each language and indeed how they are used and the particular relation that is being signalled. Essentially this means that DMs can be multifunctional and, depending on the context, convey a multitude of possible relations (Cartoni et al., 2013) leading to potential ambiguity. To help overcome this, the COMTIS project has proposed using automatic disambiguation of connectives by pre-processing occurrences and subsequently tagging the meaning. The SMT system can then learn how to translate the connectives applying what it has learnt to translate a new sentence (Cartoni et al., 2013) – it is worth noting that algorithms are first trained on manually annotated data (Popescu-Belis et al., 2012). One of the most well known resources containing sense annotation (as discussed) is the PDTB. However, it has been observed that it is difficult to consistently annotate fine-grained distinctions and in many cases inter-annotator agreement is below 70 % (Cartoni et al., 2013) which may not be precise enough for a number of NLP applications. Similarly automatic translation spotting (Véronis and Langlais, 2000) has been shown to be unreliable (around 60 %) with regard to discourse connectives (Danlos and Roze, 2011).

However, manual translation spotting and inter-changeability tests (clustering items that share the same meaning) performed by the COMTIS project have proved more reliable with some advantages over sense annotation (Cartoni et al., 2013). If a target language doesn't have a specific disambiguation then a cluster of translations is created using the agreed inter-changeability criteria, and a suitable connective (or an acceptable variant) is in theory more likely to be chosen. This process is presented as potentially being more reliable as the inter-changeability tests are performed by native speakers (Cartoni et al., 2013), although the initial set up appears to require a great deal of manual analysis. This is an interesting approach that has thus far (in terms of the COMTIS project) only been tested between English and French.

Overall using sense labelled connectives as an extra input for SMT systems showed no significant improvements, perhaps as a result of current mainstream SMT systems still trying to produce good translations of smaller sentence fragments. However, researchers have now started addressing lexical cohesion in SMT (Xiao et al., 2011; Wong and Kit,

2012; Xiong et al., 2013b) without putting as much focus on grammatical cohesion (Tu et al., 2014) although study of cohesion in SMT is still relatively limited (Xiong et al., 2013a).

Lexical cohesion is determined by identifying lexical items that form links between sentences in text (also lexical chains). A number of models have been proposed in order to try to capture document-wide lexical cohesion including: a direct reward model, a conditional probability model and an information trigger model (Xiong et al., 2013a). The direct reward model essentially rewards translations that incorporate the same word (or variant synonyms) across more than one sentence at the document level. The conditional probability model extends the reward model by measuring the appropriateness of identified cohesion by examining the likelihood that a lexical item is used across sentences in the correct manner. That is, overuse of the same lexical cohesion items may be incorrect or inappropriate and consequently affect readability (Wong and Kit, 2012).

The mutual information trigger model extends the previous models even further and takes into account lexical cohesion items that appear in successive sentences highlighting them as trigger pairs (Xiong et al., 2013a). The model is then built to determine the relationships between trigger pairs.

When the three models (incorporated into a HPBT system) were tested on NIST Chinese-English translation tasks they showed significant improvements over the baseline with the mutual information trigger model performing the best (+0.92 BLEU points) (Xiong et al., 2013a).

To achieve improved grammatical cohesion Tu et al., (2014) propose creating a model that generates transitional expressions through using complex sentence structure-based translation rules alongside a generative transfer model. The models are then incorporated into a hierarchical phrase-based translation system. The test results presented by Tu et al., (2014) show significant improvements in the process, leading to smoother and more cohesive translations. One of the key reasons for this is through reserving cohesive information during the training process by converting source sentences into "tagged flattened complex sentence structures" (Tu et al., 2014) and then performing word alignments using

the translation rules.

Work on cohesion has tended to revolve around either 'lexical cohesion' (focusing on key target words) or to a lesser extent 'grammatical cohesion'. It is argued that connecting complex sentence structures with transitional expressions is similar to the human translation process (Tu et al., 2014). Therefore improvements have been made (on a test set) showing the effectiveness of maintaining cohesive information, which includes discourse relations that provide the logical organisation of discourse segments and signal progression through a given piece of discourse.

4.2.4 Translation of Implicit Discourse Relations

It is often assumed that the discourse information captured by the lexical chains is mainly explicit. However, these relations can also be implicitly signalled in text, especially for languages such as Chinese (Yung, 2014). Yung (2014) explores DM annotation schemes such as the CDTB (4.2.2) and observes that explicit relations are identified with an accuracy of up to 94%, whereas with implicit relations this can drop as low as 20% (Yung, 2014). To overcome this, Yung proposes implementing a discourse-relation-aware SMT system, that can serve as a basis for producing a discourse-structure-aware, document-level MT system. The proposed system will use DC annotated parallel corpora, that enables the integration of discourse knowledge. Yung argues that in Chinese a segment separated by punctuation is considered to be an elementary discourse unit (EDU) and that a running Chinese sentence can contain many such segments. However, the sentence would still be translated into one single English sentence, separated by ungrammatical commas and with a distinct lack of connectives. The connectives are usually explicitly required for the English to make sense, but can remain implicit in the Chinese (Yung, 2014).

4.2.5 Anaphora Translation

Current studies in Chinese connectives have shown a number of issues, not least problems or difficulties with translating pronouns (pronominal anaphora). Anaphora translation

is a discourse-level problem, which has been studied to a degree, but is by no means solved (Hardmeier, 2012). The usage and spread of pronouns varies according to language (Russo, 2011) and so is a certain translation difficulty.

Discourse-level phenomena in SMT has been studied from a number of perspectives including domain adaptation and disambiguation. However, more recently pronominal anaphora and discourse connectives have received increased attention (Hardmeier, 2012), but as of yet exploiting such features has proved difficult, despite being a natural and obvious inclusion of discourse-level information used during human translation.

4.3 Implicitation, Explicitation, and Empty Categories

In this section we outline some approaches that have been used to deal with the topics of implicitation, explicitation, and empty categories in the context of MT. These are topics that have generated increasing interest for a number of languages in recent years.

In order to contend with implicit language phenomena special empty category tokens have been used in the Penn Treebank (Bies et al., 2002) and its extension, the Chinese Treebank (Xue and Xia, 2000). An empty category is an element that does not have a mapping to a surface word in a parse tree. Essentially, when translating such elements into the target, where they are explicitly required, it is problematic because the implicit information has to be retained, recovered, and realised from what otherwise appears to be non-existent components in the source.

In Meyer and Webber (2013) implicitation of DMs in MT is explored through a detailed corpus analysis. The work highlights how DMs in the source text are not always translated to comparable words in the target language. Disparities in how often this phenomenon occurs in human translated texts (18%) for English, French, and German as opposed to machine translated ones (8%) are observed and the work aims to more widely capture the natural implicitation of DMs in SMT.

More specifically to Chinese, Chung and Gildea (2010) examine the effects that empty categories have on MT with a specific focus on dropped pronouns (little *pro*) and control constructions (big *PRO*) . The work shows that building machine translation sys-

tems with explicitly inserted empty elements, either manually or automatically, in the training data improves the overall translation quality. They use and compare three different approaches to recover empty or null elements: pattern matching; parsing; and prediction models. Of the three, the prediction model performed the best. However, they acknowledge that there is a lot of room for improvement in order to better recover empty categories.

In Yang and Xue (2010) the term 'chasing the ghost' is used to signify the hunt for empty categories. Identification of empty categories is turned into a tagging task. Essentially, each word in a sentence is given a tag indicating whether or not it follows an empty category. A maximum entropy model is employed for the prediction of the tags. No distinctions are made between the types of tags that are identified. The results show a robust, but not breathtaking, 63% accuracy rate in recovering empty tags when an automatic parser is used as input.

Luo and Zhao (2011) also try to predict where empty categories may appear in Chinese sentences by using a statistical tree annotator supplemented with additional information. They apply the annotator to a few distinct tasks including: predicting function tags and predicting null elements. The results show favourable comparisons with previously published results using the same data. However, the results for predicting function tags and empty elements in the Chinese were obtained using human annotated data rather than automatically generated data. In addition, some of the empty categories are placed into a single position in the tree, which prevents them from being uniquely recoverable.

Instead of 'chasing the ghost' Xiang et al. (2013) outline work that 'enlists the ghost'. They use a maximum entropy model with additional syntactic features to recover empty categories and then incorporate them into a Chinese-English MT task. The results show that the recovered empty elements contribute to improvements in both word alignment tasks and the overall quality of their MT system output.

In Chapter 5, as per Steele and Specia (2014), we discuss divergences in the usage of DMs for Chinese and English. Illustrating how DMs are vital contextual links, and through a detailed corpus analysis highlight significant divergences in their usage. The

findings show how contextual omissions (implicit data) cause problems for MT systems and often lead to incoherent automatic translations.

In Chapter 6 Steele (2015) shows work that builds upon the findings in Chapter 5 with a focus on word alignments for four specific elements: 'if', 'then','because','but'. Automatic alignments are used to ascertain the occurrence of implicit markers, which is found to be quite significant. Experiments show that when artificial tokens are inserted into the data, as a proxy for these markers, and the MT systems are rebuilt, there is a significant improvement over the baseline. However, to achieve the improvement the insertions of the markers were carried out using reference data.

Clearly there is some overlap between the terms 'empty categories' and 'implicit elements', but for this work we use the latter to refer to, amongst other things, those elements with no corresponding word alignments. Our work is more general as compared to previous work and is not restricted to big or little *pro* categories, and does not rely on treebank annotations, nor on parsing.

4.4 Main Tools Used for This Research

Throughout the period of research a number of tools were used for data cleansing and model creation, as well as evaluating and visualising output.

The important tools to mention here are:

- CDEC (Dyer et al., 2010). This was the main tool used as a decoder, alignment, and learning framework for translation models. Any translation models that were built for this research were done so using CDEC.

- Fast-Align (Dyer et al., 2013). This was the standard word alignment software used with CDEC, although it was used independently as well. All the alignment information presented in this thesis was created using Fast-Align.

- Stanford Parser (Levy and Manning, 2003) (Chang et al., 2009b). This tool was used for parsing unedited Chinese text/corpora into its most useful constituent parts.

For example, white space is normally absent between Chinese characters or character groups, and so the parser helps with this in order to create the most likely tokenisation of Chinese text, which then aids the alignment process.

- CRF-Suite (Okazaki, 2007). This was used predominantly in Chapter 6 as a fast implementation of a Conditional Random Field.

- WA-Continuum. This is a self developed software tool used throughout the research to visualise word alignments and is covered in Chapter 7.

- Vis-Eval (metric viewer). Another self created software tool that was designed to facilitate the visualisation of model output translations, and a number of associated metric evaluations. This is also covered in Chapter 7 (Section 7.2).

- Bilingual Evaluation Understudy Score (BLEU). It is worth noting that BLEU (Papineni et al., 2002) is used as the primary evaluation metric to evaluate all translation output throughout this research. Other metrics were used for comparison purposes (as shown in Section 7.2), but ultimately any scores shown can be assumed to be the normal BLEU scores.

  BLEU, whilst containing known flaws is ubiquitous as it is fast to use, low cost, widely understood, and generally considered to correlate well with human judgement.

  Under the hood, BLEU essentially uses an enhanced form of precision for each sentence, taking into account uni-grams and n-grams (up to 4-grams). The larger the known n-gram (as measured against the reference) that appears in the sentence, the more fluent the output sentence is deemed to be.

  A (precision) score between 0 - 1 is given to each sentence for a whole corpus (in the output) and all scores are then combined and aggregated using the geometric mean. As BLEU tends to favour short candidate translations a brevity penalty is applied to make sure the 'accurate' shorter translations do not provide too much weight for the overall final corpus level score.

Chapter 5

5 Divergences in the Usage of Discourse Markers between English and Chinese

This section examines the divergences in DM usage across English and Mandarin Chinese. We highlight important structural differences in composite sentences extracted from a number of parallel corpora, and show examples of how these cases are dealt with by popular SMT systems. Numerous significant divergences, such as contextual omissions, were observed, which can lead to incoherent automatic translations.

5.1 Discourse

In general "discourse" is used to signify an arbitrary length of coherent language-based communication consisting of either phrases, sentences or utterances (Zuffery and Degand, 2013). With respect to NLP, and more specifically, SMT, discourse is mainly concerned with both written text and spoken dialogue consisting of some connected sequential units.

On a fundamental level discourse is linked in a meaningful way (lexical cohesion) by DMs (also known as discourse connectives), which separate the discourse into discourse segments or language structures, such as words, phrases, clauses or composite sentences (Tsou et al., 1999), each of which contains a local coherence and context. However, DMs cover a range of connectives, conjunctions, conjunctives and other cue words (see Chapter 3) and can be difficult to define precisely.

Here we examine the usage of a set of frequently used DMs in Chinese[17] and English, highlighting some natural and common divergences observed in parallel corpora, and some of the problems that arise when the contextual information that surrounds them is not utilised by SMT systems. The focus is on Chinese into English translations. The results were produced from inspecting four corpora of various genres, domains and sizes, comparing given DMs in Chinese sentences against DMs in the English parallel human translation. Only DMs within sentences (intra-sentence DMs), rather than across discourse segments, were used for the analysis. The study shows that the parallelism in the usage of DMs in the two languages varies significantly across corpora. It also shows substantial divergences in the usage of DMs in a large proportion of cases. This evidences the problem of using such parallel corpora as a source of information to build SMT systems without special treatment of DMs.

Popular online SMT systems were also used to translate the Chinese, with the resulting automated translation being compared to the given human translation, hence illustrating their limitations. The results show that these SMT systems are often unable to deal with the complex changes in word order and, because of DMs, struggle with contextual omissions, even across closely linked sentential clauses. As the sentences become more complex the problems are further compounded and more errors occur in the automated translations, ultimately suggesting that too much information is lost when the context carried by DMs is not utilised by SMT systems.

5.2 Discourse Markers in Chinese

Chinese and English stem from two very different language families (Sino-Tibetan and Indo-European respectively) which can be a chief cause of translation difficulty (Chang et al., 2009a). For example, Chinese is logographic and does not use inflection, relying on generating meaning through word order, which can often be quite flexible. Moreover, the positioning and order of connective markers is very fluid and syntactically, markers can take many positions. English, on the other hand, has an alphabet and uses a degree of

[17] Mandarin Chinese, the main standardised language of China (Swan and Smith, 2004).

inflection with a relatively fixed word order (Li, 2008).

Defining DMs is not necessarily a trivial task. Chinese uses a rich array of DMs to create links in both simple and complex sentences (Tsou et al., 1999). Chinese conjunctions appear in two main types: those linking words or short phrases (simple conjunctions) such as: 和(hé - and), 跟(gēn - and/with), 或(huò - or) as in 刀和叉(dāo hé chā - knife and fork), and those that link clauses (composite conjunctions). Conjunctives are also used, often appearing in the main (usually second) clause of a sentence and link back to the previous clause (Po-Ching and Rimmington, 2004). Additionally, there are instances where clauses may be linked in a sentence without the use of any DM (zero connective structures). In these cases the meaning or context is strongly inferred across the clauses, leading to the creation of sentences that have natural omissions, which can cause problems for MT systems.

5.3 Settings: Corpora and SMT Systems

We used four well known corpora (as listed in Section 4.1.1) for gathering the data necessary for observing DM frequency, usage, and any pertinent translations:

- Basic Travel Expression Corpus (BTEC)
- Foreign Broadcast Information Service (FBIS) Corpus
- Ted Talks Corpus (TED)
- Multi-UN Corpus (UN)

At the time of this study the SMT systems used to produce the automatic translations are Google Translate[18] and Bing Translator[19]. Whilst these are specific commercial translation tools and they may not represent the best quality translation systems for Chinese-English, they are good representatives of statistical translation approaches (now they use NMT methodology), known to use state of the art techniques and achieve reasonable

[18]http://translate.google.com
[19]http://www.bing.com/translator

translation quality. In addition they are freely available, making it possible to reproduce and expand the analysis presented here.

5.4 Analysis of Chinese Discourse Markers

In this section we examine the main types of Chinese DMs, including conjunctions for composite sentences, sequential paired conjunctions and zero connectives (Hutchinson, 2004; Po-Ching and Rimmington, 1998, 2004, 2010; Ross and Sheng Ma, 2006; Teachers, 2010). Our first step was a simple quantitative analysis to identify the most commonly used DMs in our corpora, so that we could select a few cases of interest to analyse in more detail. Table 7 shows the proportion of sentences containing the ten most frequent DMs in the four different corpora. It also shows one or more frequent English translations for each DM, but we note that variants of these translations are possible.

Table 7: Ten most frequently occurring DMs in the four corpora.

TED	*UN*
因为(4.72%) : because	因此(1.70%) : so/therefore
如果(4.32%) : if	以便(1.42%) : so that
所以(4.05%) : so/therefore	因为(1.24%) : because
但是(3.58%) : but	由于(1.22%) : due to/as a result of
或者(1.68%) : or	如果(1.05%) : if
还有(1.59%) : furthermore	而且(1.04%) : moreover
那么(1.59%) : then/in that case	为了(0.88%) : in order to
而且(1.47%) : moreover	但是(0.81%) : but
并且(1.34%) : and also	并且(0.73%) : and also
因此(1.24%) : so/therefore	虽然(0.62%) : although
FBIS	*BTEC*
因为(1.39%) : because	如果(1.18%) : if
如果(1.30%) : if	但是(1.10%) : but
因此(1.19%) : so/therefore	那么(0.44%) : then/in that case
为了(1.13%) : in order to	还是(0.39%) : or
由于(1.10%) : due to/as a result of	所以(0.29%) : so/therefore
但是(1.01%) : but	因为(0.25%) : because
而且(0.85%) : moreover	或者(0.23%) : or
虽然(0.80%) : although	并且(0.17%) : and also
然而(0.79%) : however/but	只有(0.17%) : only
甚至(0.72%) : even	而且(0.13%) : moreover

While the percentages of sentences containing specific DMs in Table 7 may seem

small at first, overall DMs are present in a significant proportion of sentences. The frequency analysis highlights certain trends, for instance 如果(rúguǒ – if) and 因为(yīnwèi – because) have a relatively high frequency in all four corpora. 因为(yīnwèi) is classed as one of the high frequency (causal) connectives (Wang and Huang, 2006) and is considered to have a strong correlation in usage with 'because'. In what follows we pinpoint some of the divergences in the use of these markers through examples of constructions, and connect these divergences to the behaviour of SMT systems when faced with such constructions.

Example 11 shows the 因为(yīnwèi) DM being used in a relatively short causal sentence, and it is clear that the SMT system has problems with the DM, dropping it completely from its position before the comma.

(11) 他因为病了，没来上课。[20]

he because ill, not come class.

Because he was sick, he didn't come to class. (Ross and Sheng Ma, 2006)

He is ill, absent. (Bing)

In Example 11 the two parts of the sentence appear to have a very weak link in the translation as the DM is simply not used at all in the automated translation. The information after the comma (in the Chinese sentence) is correct and as Chinese does not use inflection, a sentence segment similar to 'did not come to class' should appear in the translation rather than simply having 'absent'.

In Example 12 the problem seems to be the reverse. The 因为(yīnwèi – because) being present in the Chinese sentence causes problems for the SMT system as it tries to force 'because' into the translation (rather than omitting it) and by doing so significantly alters the meaning.

[20]Each example in this section has the following format: Line 1 is the correct Chinese in characters; line 2 is a literal word-for-word translation; line 3 is the given translation and line 4 is (usually) the best translation returned by the SMT system. In some cases more than one automated translation is given for comparison purposes.

(12) 你因为这个在吃什么药吗?

you because this (be) eat what medicine [MA]

Have you been taking anything for this? (BTEC)

What are you eating because of this medicine? (Google)

The automated translation gives the impression that the person has changed their diet due to having medicine, rather than their being required to take medicine for an ailment.

5.4.1 Sequential Constructions: Paired Conjunctions/Conjunctives

Paired DMs are frequently used in Chinese (Xue, 2005) and feature in many translations of complex sentences. Some paired constructions are formed using two conjunctions, but other formations are also possible such as: 'conjunction. . . conjunctive'. Typical conjunctives include: 才(cái - only/only if/ not unless), 就(jiù - then/that), 却(què - but/yet/while) and are commonly treated as connecting referential adverbs (Po-Ching and Rimmington, 2004) . Conjunctions tend to appear in both clauses, while conjunctives frequently appear in just the second clause. They represent even more challenging problems for both human and machine translation.

Table 8 shows (for each corpus) the proportion of sentences that contain at least one occurrence of the given paired marker patterns. The main outcome of this frequency analysis is that for each corpus the ...一...就... (. . . yī. . . jiù. . .) pattern appears with the highest frequency. However, manual inspection of a random sample of sentences showed that the ...一...就... (. . . yī. . . jiù. . .) structure was only being used as a sequential paired marker construction in around one quarter of the cases.

Chinese does not have a specific word which maps one-to-one exactly with 'then' and so 就(jiù) and 那么(nàme - so) are often utilised to perform a similar function (Ross and Sheng Ma, 2006) . It is difficult to categorise 就(jiù) on its own as it serves numerous functions. Many other characters such as 来(lái) and 的(de) can also be difficult to categorise for a similar reason, but perhaps none more so than the character '一' (yī - one/single/ whole/same. . .) which covers six pages in the Oxford Chinese dictionary.

Table 8: Ten most frequently occurring paired DMs in the four corpora.

TED	*UN*
...一...就... (3.67%) : once/as soon as, (then)	...一...就... (0.92%) : once/as soon as, (then)
...如果...就... (1.33%) : if,(then)	...越...越... (0.30%) : more, more
...如果...那... (0.95%) : if, (then)	...由于...因... (0.24%) : due to, because
...也...也... (0.49%) : also, and	...如果...就... (0.22%) : if,(then)
...越...越... (0.49%) : more, more	...不仅...而且... (0.21%) : not only, but also
...从...开始... (0.48%) : starting from...	...从...起... (0.17%) : starting from...
...是...还是... (0.48%) : [be], or	...从...开始... (0.14%) : starting from...
...如果...那么... (0.34%) : if, (then)	...是...还是... (0.14%) : [be], or
...不是...而是... (0.29%) : not, but(is)	...虽然...但是... (0.12%) : although, but
...从...起... (0.27%) : starting from...	...也...也... (0.11%) : also, and
FBIS	*BTEC*
...一...就... (2.20%) : once/as soon as, (then)	...一...就... (0.28%) : once/as soon as, (then)
...越...越... (0.63%) : more, more	...如果...就... (0.22%) : if, (then)
...也...也... (0.40%) : also, and	...从...开始...(0.15%) : starting from...
...从...起... (0.38%) : starting from...	...如果...那... (0.10%) : if, (then)
...如果...就... (0.36%) : if,(then)	...从...起... (0.09%) : starting from...
...从...开始... (0.35%) : starting from...	...是...还是... (0.06%) : [be], or
...不仅...而且... (0.30%) : not only, but also	...只要...就... (0.06%) : as long as, (then)
...是...还是... (0.27%) : [be], or	...又...又...(0.05%) : both, and
...既...又... (0.25%) : both, also	...越...越... (0.03%) : more, more
...既...也... (0.24%) : both, also	...的话...就... (0.03%) : ...if, (then)

By themselves 一(yī) and 就(jiù) can be ambiguous, but as a sequential construction they work together as a pair in a specific pattern with a relatively fixed meaning. Example 13 shows a short five-character sentence that uses the ...一...就... (...yī...jiù...) pattern as a sequential paired construction to mean: '...no sooner...than...'; 'the moment...'; 'as soon as...'; 'once...'

(13) 他一学就会。

he as soon as study then can.

He learned it (the trick) in a jiffy. (Manser, 2009)

He learn. (Google)

In Example 13 it is clear that very little concrete information can be extracted from the

five characters alone, and there is a lot of inference such as the speed in which the person learned to do something (in this case - a trick). To identify both the 'trick' and 'speed' would require additional contextual information.

The overarching pattern for the ...一...就... (...yī...jiù...) construct is fairly simple: ...一VP^{a} 就 VP^{b}

The 一(...yī...) should come immediately before the prepositional phrase and/or verb or verb phrase (Ross and Sheng Ma, 2006), although it can have some subject information that precedes it. In the case of Example 13 a pronoun is used for the subject.

It is possible that by itself the sentence in Example 13, while grammatically correct, has too much inference for an SMT system to manage and sentences that contain more information may produce better translations.

The actions in the structure do not have to be related and the subjects in each clause do not have to be the same, but it is often the case that the second action is as a direct result of the first.

(14) 一有空位我们就给你打电话。

As soon as have space we then give you make phone.

We'll call you as soon as there is an opening. (BTEC)

A space that we have to give you a call. (Google)

In Example 14 the SMT system tries to remain closer to the actual order of the given sentence, but once again misses the 'as soon as'. If the word order is to be kept close to the original then a sentence similar to 'as soon as we have a vacancy (then) we will give you a call' could be used.

5.4.2 Linking Clauses Without Discourse Markers (Zero Connectives)

The zero connective (Po-Ching and Rimmington, 2004) is often used to link closely set clauses where the meaning of the second clause is contextually implied by the meaning of the first clause. This can be done through repetition, answering, or qualifying conditions as in Example 15, or for rhythmic balance (Po-Ching and Rimmington, 2010).

(15) 东西太贵, 我不买。

things too expensive, I not buy

If things are too expensive, I won't buy them. (Po-Ching and Rimmington, 2010)

Too expensive, I do not buy it. (Google)

Having a zero connective is perhaps the ultimate contextual omission. In this case, the SMT system appears to translate the Chinese word for word rather than actually applying meaning. The gist of the condition is evident, but the translation is not adequate. Manual insertion of two standard DMs (akin to '**if**' and '**then**') into the sentence is actually required for the SMT system to produce a better output as shown in Example 16.

(16) **如果** 东西太贵，我**就** 不买(了)。

If things too expensive, I then not buy(le).

If something is too expensive, I do not buy it. (Google)

5.5 Analysis of Chinese and English discourse markers in parallel corpora

In this Section we perform a quantitative analysis on the usage of DMs in both Chinese and English (human translation). SMT systems learn translation models primarily from parallel corpora with examples of translations aligned at the sentence level. The goal of this analysis is to study whether Chinese markers and their corresponding English markers appear in sentences that are aligned in parallel corpora. For a given DM, a high percentage of aligned sentences containing the marker in both Chinese and English could be an indication that learning the translation of such a marker from the corpus is potentially feasible. On the other hand, a low percentage of aligned sentences containing both Chinese and English markers could be an indication that the markers might be dropped or translated using different linguistic constructs, making the learning of SMT models a more difficult task.

Given that we start the analysis with Chinese DMs, a question that arises is how to find their corresponding English DMs. Each of the given DMs (Tables 7 and 8) is relatively

Table 9: Frequencies of six Chinese DMs and their corresponding translations in parallel corpora.

Chinese Marker	Occurrence rate in Chinese (%)				Occurrence rate in human translation (%)				Appear in both the Chinese and English translation (%)			
	BTE	FBIS	UN	TED	BTE	FBIS	UN	TED	BTE	FBIS	UN	TED
因为(because)	0.25	1.39	1.24	4.72	0.20	1.01	0.48	3.92	80	73	39	83
如果(if)	1.18	1.30	1.05	4.32	1.15	1.09	0.76	3.84	89	84	72	89
因此(consequently)	0.02	1.19	1.70	1.24	0.02	0.83	1.09	1.07	100	70	64	86
但是(but)	1.10	1.01	0.81	3.58	1.07	0.89	0.54	3.19	97	88	67	89
而且(moreover)	0.13	0.85	1.04	1.47	0.13	0.59	0.69	1.15	100	69	66	78
虽然(although)	0.02	0.80	0.16	0.36	0.02	0.65	0.15	0.15	100	81	94	42

common, but can naturally have variance in the associated translations. For example, a strong link has already been suggested between 因为(yīnwèi) and 'because', but there are numerous comparable ways of uttering or writing 'because' such as: 'in light of', 'for this reason', 'as a result of' (Macmillan) (Roget's Thesauraus). For this section, interchangeable values are classed as variance rather than ambiguity. Ambiguity is taken to mean a word that has numerous different functions as per the individual characters '一' yī and '就' jiù as discussed.

Table 9 shows the occurrence percentages of six frequently used Chinese DMs in the four corpora. The first column shows the Chinese DM with its commonly associated English equivalent. Column two shows the occurrence rate of the Chinese marker in sentences across the corpora. Column three shows the occurrence rate where a directly equivalent English DM (with variance included) is used in the parallel translations (e.g. 因为= 'because' or a variant of 'because'); that is, for each set of sentences with a given Chinese DM, a subset is formed from the parallel translations of the sentences. The percentages in column three show the size of the resulting subsets compared to the size of the whole corpus. The final column shows the percentages of sentences that contain, within a set, both the Chinese DM along with the equivalent usage of an English DM in the translation. The percentages in the fourth column can be used as general measure of

the strength of correlation.[21]

We note that the source language of our corpora is not always Chinese. For TED it is English, while for UN it could be any of the six languages. BTEC and FBIS however consist of segments originally in Chinese, and their translations into English. Therefore the implications of the numbers in Table 9 will be different for different corpora. For example, with the source being English for the TED corpus, the tokens in question are more likely to be explicitly contained within the given sentences. This means, in turn, the parallel Chinese sentences, which are created during translation, are more likely, where it makes sense, to contain a higher degree of explicitation than if the situation were reversed. That is English-Chinese means more instances of DMs and Chinese-English means fewer.

Overall, the numbers show that in short everyday sentences (BTEC) the main DMs are used as expected (e.g. 因为 maps closely to 'because'). As the sentences become more complex and are used at a higher level (FBIS and TED), then the way DMs are used becomes more fluid. The markers appear to be increasingly omitted or absorbed into the general meaning of a clause rather than translated directly. As expected with the UN corpus, where complex language is used and discourse is divided into subsections, addenda, and annexes, there is even less need for certain markers and there are inevitably fewer occurrences of items such as 'if' and 'but'.

(17) 这将是一次规模盛大, 而且受到广泛国际关注的聚会.

This will be one scale grand, moreover receive wide international attention [DE] meeting.

This will be a grand gathering with wide international concern. (FBIS)

This will be a grand scale, but widespread international concern gatherings. (Google)

In Example 17, the 而且(érqiě - moreover) is serving as a link that brings together the qualities of the meeting; that is, it will be on a 'grand scale' and will receive 'wide

[21]It must be noted that whilst the percentages show trends, there is still a small degree of error where less common variant phrases may have possibly been used in the parallel translation (e.g. because = this is down to). Detailed discussion of further variance is beyond the scope of this section and can be considered in future work. The given percentages are considered to offer a close enough approximation for the related discussion.

international attention'. Clearly the human translation is very succinct and does away with the need for the 'moreover' or 'furthermore' type link.

For an SMT system to reach a similar translation it would need to be aware of when to drop the marker, and how to reorder the sentence accordingly. Additionally the character 的[DE], a structural particle often used for modifying nouns, adds complication as grammatically it implies that the described qualities (scale and attention) belong to the meeting, which is not necessarily an easy connection to automatically recognise.

5.6 Conclusion

Chinese and English both belong to very different language families leading to numerous structural differences between the two languages including differing word order and the use of DMs. DMs in particular provide a level of lexical cohesion between phrases and clauses, but are not always utilised adequately during the automated translation process. This means that sentential positioning is often incorrect and certain words are frequently omitted leading to unclear translations with a loss of context and information.

In many cases Chinese discourse has significant subject inference carried across clauses and sentences leading to contextual omission of many items (often pronouns) within a sentence. Example 18 shows a modified version of Example 11 where the pronoun and second marker have been manually inserted into the Chinese sentence. With the extra information Bing returns a better translation, highlighting the importance of preserving DMs and contextual information.

(18) 他因为病了, 所以他没来上课。(modified version of Example 11)

he because ill, so he not come class. (extra 'he' and 'so' in the 2nd clause)

Because he was sick, he didn't come to class.

He is ill, so he did not come to class. (Bing)

In the case of paired DMs, especially with the 一(yī) and 就(jiu) structure, the SMT systems struggled with inference and disambiguation, often failing to spot the 'as soon as' relation.

Chapter 6

6 Improving Translation of Discourse Connectives and Discourse Relations in SMT

We have established that DMs are essential, widely used cohesive devices that connect segments of utterances or written text. We have also established that across different languages there is much divergence in their usage, placement, and frequency (Chapter 5). Ultimately this is still considered to be a major problem for MT and for SMT is particular.

Here we present an examination of the difficulties around translating DMs in SMT and we also detail aspects of modelling cohesive devices within sentences. Initial experiments showed promising results for building a prediction model that uses linguistically inspired features to help improve word alignments with respect to the implicit use of cohesive devices, which in turn leads to improved translations. Part 6.1.1 outlines the initial motivation and research as seen in the preliminary corpus analysis (Chapter 5). It covers examples that highlight various problems with the translation of (implicit) DMs, leading to an initial intuition. Section 6.1.2 looks at experiments and word alignment issues following a deeper corpus analysis and discusses how the intuition led towards developing the methodology used to study and improve word alignments. It also includes the results of the experiments that show positive gains in BLEU. Section 6.1.3 provides an outline of future considerations.

Building on the items detailed in Section 6.1 we create a prediction model that aims to predict where we can insert a representatives token to improve word alignments and

allow the MT system to leverage the extra information. The entire methodology of this process is discussed in Section 6.2 with the subsequent related experiments being detailed in Section 6.3.

Automatic Evaluation It is worth noting that BLEU is our main evaluation metric used to evaluate MT output quality. Although it lacks the specific ability to evaluate attempts to address a given linguistic phenomenon correctly, it is a widely accepted evaluation tool. Also an improvement in the BLEU score should be representative of improved translations as a whole. In addition, as automatic evaluation will not capture all the linguistic phenomena that we wish to examine, it is anticipated that a certain level of manual inspection and evaluation is required hence our development of significant visualisation tools (Chapter 7).

6.1 Modelling Discourse Markers in Chinese Sentences

In Chapter 5 we presented an overview of our initial corpus analysis, which led to our finding useful motivating examples that showed deficiencies in how SMT systems translate DMs. Here we build upon that work looking more deeply into an extended corpus analysis and ways in which we can model DMs in Chinese.

6.1.1 Motivation

This section draws attention to deficiencies caused from under-utilising discourse information and examines divergences in the usage of DMs. The final part of this section outlines the intuition gained from the given examples and highlights the approach to be undertaken for modelling DMs in Chinese. For our deep corpus analysis, research, and experiments three main parallel corpora are used (listed in Section 4.1.1):

- Basic Travel Expression Corpus (BTEC)
- Foreign Broadcast Information Service (FBIS) Corpus
- Ted Talks Corpus (TED)

In Chapter 5 we showed, through an initial corpus analysis, that Chinese uses a rich array of DMs including: simple conjunctions, composite conjunctions, and zero connectives, which can cause problems for current SMT approaches.

We also showed (through Examples 11 and 12) concrete translation output that highlighted difficulties the MT systems had with DMs. For example, both of the examples showed 'because' (因为) being used in different ways. In each case the automated translations fall short. In Example 11 the dropped (implied) pronoun in the second clause is thought to be the problem, whilst in Example 12 significant reordering is needed, which is hard to capture from a single sentence. By itself the sentence appears somewhat exophoric and the meaning cannot necessarily be gleaned from the single segment of text alone. Example 14 in Chapter 5 also introduced the '一' and '就' ('as soon as') construct, and showed how SMT systems found it difficult to process.

(19) 她一回来我就让她给你打电话。

she as soon as returns I then get her give you phone call.

I'll tell her to call you as soon as she returns. (BTEC)

In Example 19 (as with Example 14) the characters '一' and '就' are working together as coordinating markers in the form: ...一VP^a 就 VP^b.

Figure 2[22] shows a visualisation of the WA for the 'as soon as' sentence given in Example 19. In the image it is immediately obvious that the Fast-Align tool failed to map the 'as soon as ... then' structure to ...一... 就... . That is, columns 7, 8, 9, which represent 'as soon as' in the English have no alignment points whatsoever. Yet, in this case, all three items should be aligned to the single element '一' which is on row 1 on the Chinese side. Additionally, the word 'returns' (column 11), which is currently aligned to '一' (row 1) should in fact be aligned to '回来' (return/come back) in row 2. This misalignment could be a direct side-effect of having no alignment for 'as soon as' in the first place. Consequently, the knock-on effect of poor word alignment, especially around markers, as in this case, will lead to the overall generation of poorer translation rules.

[22]The boxes with a '#' inside are the alignment points and each coloured block (large or small) is a minimal-biphrase. See Chapter 7.

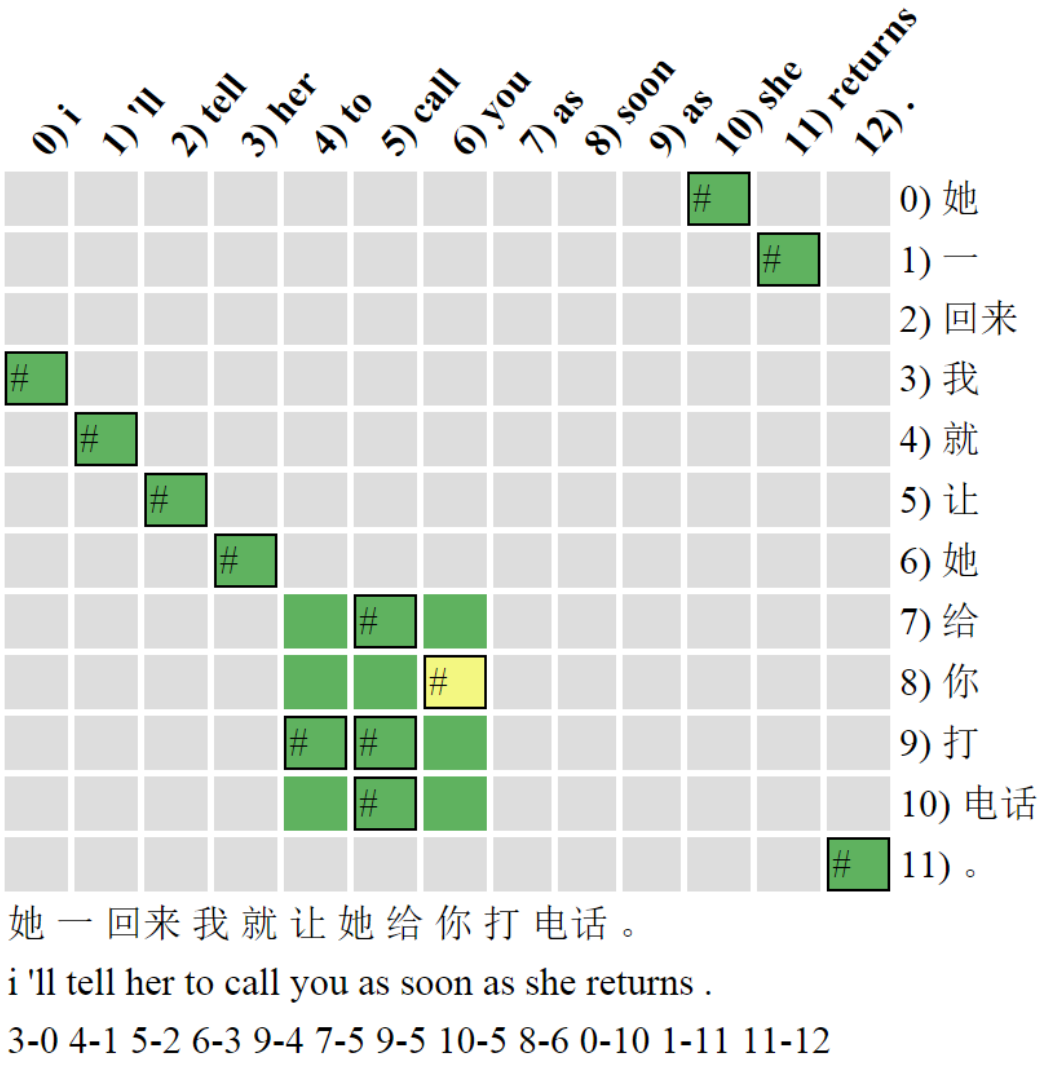

Figure 2: A visualisation of Fast-Align word alignments for the given parallel sentence (Chinese-English), showing a non-alignment of 'as soon as'.

(20) 他因为病了, 所以他没来上课。

he because ill, **so he** not come class.

Because he was sick, he didn't come to class.

He is ill, so he did not come to class. (Bing)

Example 18 in Chapter 5 (shown again here in Example 20) highlights how manually inserting an extra 'so'(所以) and 'he' (他) in the second clause of the Chinese sentence improves the translation output. Grammatically these extra characters are not required for the Chinese to make sense. However, the interesting point is that the extra information (namely 'so' and 'he') has enabled the system to produce a much better final translation.

From the given examples (both here and in Chapter 5) it appears that both implicitation and the use of specific DM structures can cause problems when generating automated translations. The highlighted issues suggest that making markers (and possibly, by extension, pronouns) explicit, more information becomes available, which can support the extraction of word alignments. Although making implicit markers explicit can seem unnatural and even unnecessary for human readers (Example 20), it does follow that if the word alignment process is made easier by this explicitation it will lead to better overall translation rules and ultimately better translation quality.

6.1.2 Experiments and Word Alignments

This section examines our research and experiments used to measure the extent of the difficulties caused by DMs. In particular the focus is on automated word alignments and problems around implicit and misaligned DMs. The work discussed in Section 6.1.1 highlighted the importance of improving word alignments, and especially how missing alignments around markers can lead to the generation of poorer rules.

Before progressing onto the experiments an initial baseline system was produced according to detailed criteria (Chiang, 2007; Saluja et al., 2014). The initial system was created using the ZH-EN data from the BTEC parallel corpus (Paul, 2009) (Section 6.1.1). Fast-Align is used to generate the word alignments and the CDEC decoder (Dyer et al., 2010) is used for rule extraction and decoding. The baseline and subsequent systems discussed here are hierarchical phrase-based systems for Chinese to English translation.

Once the alignments were obtained the next step in the methodology was to examine the misalignment information to determine the occurrence of implicit markers. A variance list was created[23] that could be used to cross-reference discourse markers with appropriate substitutable words (as per Table 5). Each DM was then examined in turn (automatically) to look at what it had been aligned to. When the explicit English marker was perceived to be aligned correctly, then no change was made. If the marker was aligned to an unsuitable word, then an artificial marker was placed into the Chinese in the nearest free space to

[23]The variance list is initially created by filtering good alignments and bad alignments by hand and using both on-line and off-line (bi-lingual) dictionaries/resources.

DM	BTEC	FBIS	TED
if	25.70%	40.75%	23.35%
then	21.00%	50.85 %	40.47%
because	23.95%	32.80%	16.48%
but	29.40%	39.90%	27.08%

Table 10: Misalignment information for the 3 corpora.

System	DEV	TST
BTEC-Dawn (baseline)	34.39	35.02
BTEC-Dawn (if)	34.60	35.03
BTEC-Dawn (then)	34.69	35.04
BTEC-Dawn (but)	34.51	35.21
BTEC-Dawn (because)	34.41	35.02
BTEC-Dawn (all)	34.53	35.46

Table 11: BLEU scores for the experimental systems.

that word. Finally, if the marker was not aligned at all then an artificial marker was inserted into the nearest free space by number[24]. A percentage of misalignments[25] across all occurrences of individual markers was also calculated.

Table 10 shows the misalignment percentages for the four given DMs across the three corpora. The average sentence length in the BTEC Corpus is eight tokens, in the FBIS corpus it is 30 tokens, and in the TED corpus it is 29 tokens. The scores show that there is a wide variance in the misalignments across the corpora, with FBIS consistently having the highest error rate.

Initially tokens were inserted for single markers at a time, but then finally with tokens for all markers inserted simultaneously. Table 11 shows the BLEU scores for all the experiments. The first few experiments showed improvements over the baseline of up to +0.30, whereas the final one showed good improvements of up to +0.44.

[24]The inserts are made according to a simple algorithm, and inspired by the examples in Section 3.

[25]A non-alignment is not necessarily a bad alignment. For example: '正反' = 'positive and negative', with no 'and' in the Chinese. In this case a non-alignment for 'and' is acceptable.

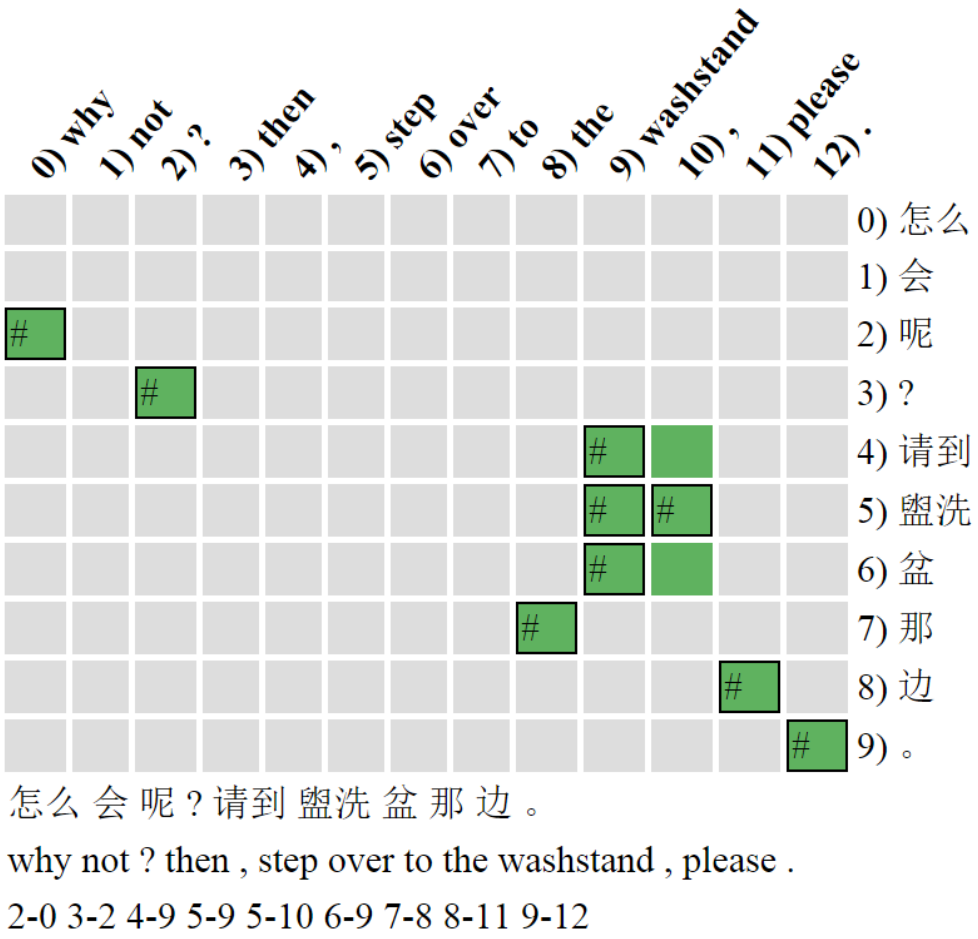

Figure 3: Visualisation of word alignments showing no alignment for ‘then’ in column 3.

After running the experiments the visualisation of a number of word alignments (as per Figures 2, 3, 4) were examined and a single example of a ‘then’ sentence was chosen at random. Figure 3 shows the word alignments for a sentence from the baseline system, and Figure 4 shows the word alignments for the same sentence, but with an artificial marker automatically inserted for the unaligned ‘then’. To achieve the latter, Fast-Align is retrained after making inserts.

The differences between the word alignments in the figures are subtle, but positive. For example, in Figure 4 more of the question to the left of ‘then’ is captured correctly. Moreover, to the right of ‘then’, ‘over’ has now been aligned quite well to ‘那边’ (over there) and ‘to’ has been aligned to ‘请到’ (please - go to). Perhaps most significantly though is that the mish-mash of alignments to ‘washstand’ in Figure 3 has now been

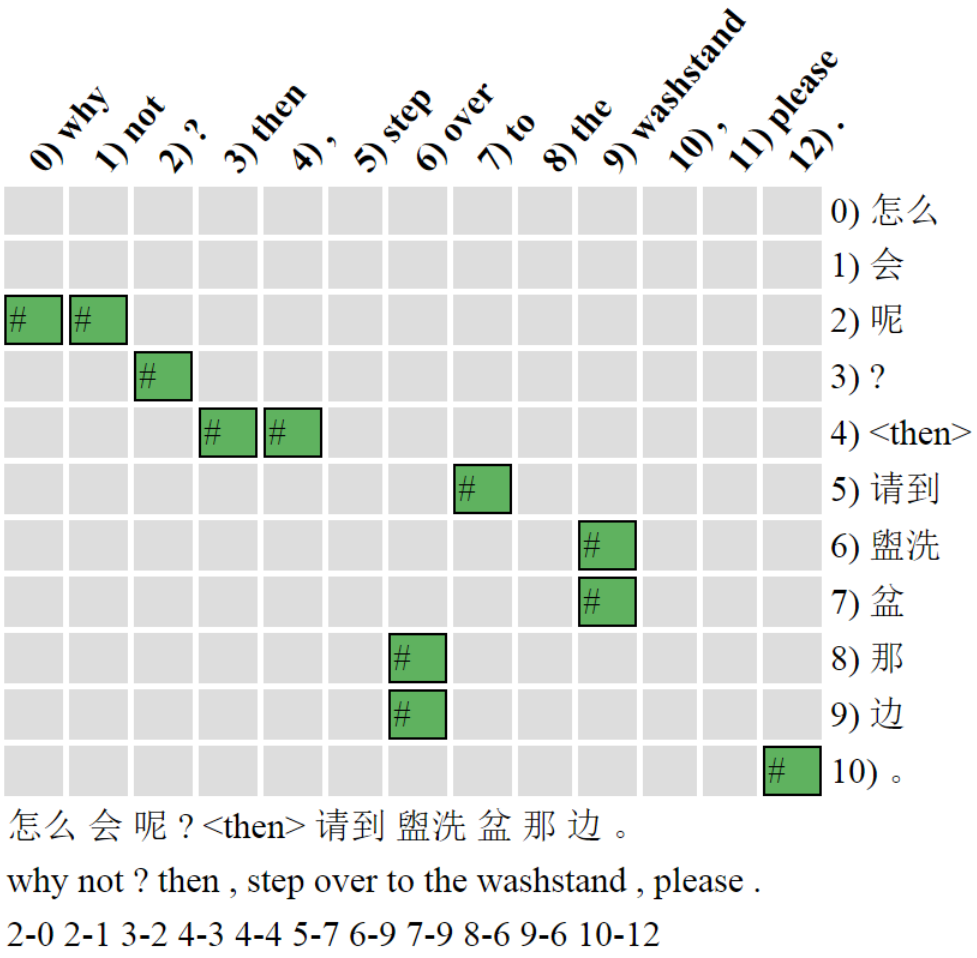

Figure 4: Visualisation of word alignments showing the artificial marker '<then>' and a smoother overall alignment.

replaced by a very good alignment to '盥洗盆' (washbasin/washstand) showing an overall smoother alignment. These preliminary findings indicate that there is plenty of scope for further positive investigation and experimentation.

6.1.3 Modelling Cohesive Devices Within Sentences

Having addressed the limitations of current SMT approaches, the focus has moved on to looking at cohesive devices at the sentential level.

Even at the sentence level there exists a local context, which produces dependencies between certain words. The cohesion information within the sentence can hold vital clues for tasks such as pronoun resolution, and so it is important to try to capture it.

Simply looking at the analysis in Section 6.1.2 provides insight into which other avenues should be explored, including:

- Expanding the number of DMs to capture complex markers (e.g. as soon as).
- Improving the variance list to capture more variant translations of marker words. It is also important here to include automated filtering for difficult DMs (e.g. cases where 'and' or 'so' are not being used as specific markers can perhaps make them more difficult to align).
- Developing better insertion algorithms to produce an improved range of insertion options, and reduce damage to existing word alignments.
- Looking at using alternative/additional evaluation metrics and tools to replace or complement BLEU. This could produce more targeted evaluation that is better at picking up on individual linguistic components such as DMs and pronouns.

However, the final aim is to work towards a true prediction model using parallel data as a source of annotation. Creating such a model can be hard mono-lingually, whereas a bilingual corpus can be used as a source of additional implicit annotation or indeed a source of additional signals for discourse relations. The prediction model should make the word alignment task easier (either through guiding the process or adding constraints), which in turn will generate better translation rules and ultimately should improve MT.

The ultimate aim is to use bilingual data as a source of additional clues for a prediction model of Chinese implicit markers, which can, for instance, guide and improve the word alignment process leading to the generation of better rules and smoother translations.

6.2 Working Toward a Prediction Model for Improving the Translation of Discourse Markers for Chinese into English

In MT implicitation can occur when elements such as DMs and pronouns are not expected or mandatory in the source language, but need to be realised in the target language for a coherent translation (as highlighted in Section 6.1). These 'implicit' elements can be seen as both a barrier to MT and an important source of information. However, identifying where such elements are needed and producing them are non-trivial tasks. Here we examine the effect of implicit elements on MT and propose methods to identify and make them explicit. As a starting point, we use human translated and aligned data to decide where to insert place holders for these elements.

We then fully automate this process by devising a prediction model to decide if and where implicit elements should occur and be made explicit.

Our experiments compare SMT models built with and without these explicitation processes. Models built on data marked for discourse elements show substantial improvements over the baseline.

6.2.1 Introduction

One of the main challenges in MT is to model the multitude of intrinsic differences that occur between the source and target languages. The problem is even more critical when considering distant language pairs such as Chinese and English. Chinese, for example, has a flexible grammar, relatively free word order (Gao, 2008), and is a prolific 'pro-drop' language (Huang, 1989). In addition, the application of cohesive devices (e.g. conjunctions) is one of the most prominent features that distinguishes Chinese and English (Wu, 2014). For instance, implicit links (i.e. the absence of explicit markers) are very common in Chinese and where a relation is not made explicit it can be inferred from context. However, in MT producing the correct explicit information on the target language when it is not required on the source language poses a significant barrier that often leads to poor quality automated translations.

In Chapter 5 and Section 6.1 we gave examples of problems caused by DMs and their

usage. In addition we showed that by inserting some 'token' markers we could improve WA information, which in turn led to smoother translations.

Here we examine some of the effects that implicit elements have on MT. We also implement methods for recovering some of the inferred information by inserting explicit place holder tokens into the source data to help inform the automatic alignment and decoding processes. We create an initial benchmark using human translations and oracle alignments (correct word alignments provided by human experts), which we then try to automate using a classifier to predict if and where to insert place holder tokens.

Our primary results show a significant improvement over the baseline models with no place holders for discourse elements (+1 in BLEU) and are close to those obtained with annotations derived from manually produced translations and alignments.

The remainder of this Chapter is organised as follows: Section 6.2.2 explains in detail how we built our benchmark corpus based on datasets translated and word-aligned by humans. Section 6.2.3 details our methods used for finding implicit elements and inserting place holder tokens into our data. We also discuss our initial work on building a prediction model. Our experiments, set-up and results are outlined and discussed in Section 6.3. Finally, Section 6.3.4 presents our conclusions and potential directions for future work.

6.2.2 A Benchmark Corpus

Here we describe the pre-processing of a benchmark corpus using human translated and aligned data, which we then use to build and evaluate approaches to make discourse elements explicit.

The Data The data used to build our benchmark corpus came from sections of the Gale Project provided by the Linguistic Data Consortium (LDC)[26] catalogue[27]. Each section consists of manually translated sentences from news and web broadcasts and contains oracle (i.e. produced by expert linguists) word alignments, as well as additional annotations signalling items such as non translated elements and other metadata.

[26]https://catalog.ldc.upenn.edu/

[27]LDC2012T16, LDC2012T20, LDC2012T24, LDC2013T05, LDC2013T23, LDC2014T25, LDC2015T04, LDC2015T18, LDC2015T06

Our final corpus is made up of a total of 43693 usable parallel aligned sentences[28] consisting of approximately 1.23M English words and 955K Chinese words:

- GALE Chinese-English Word Alignment and Tagging – Broadcast Training Parts 1-4. Total = 19621 usable sentences.
- GALE Chinese-English Word Alignment and Tagging Training – Newswire and Web Parts 1-4. Total = 17966 usable sentences.
- GALE Chinese-English Parallel Aligned Treebank – Training. Total = 6106 usable sentences.

Building the Sentences Example 21 shows a typical sentence in its original format. The Chinese is character segmented and the English is space delimited. The word alignments reflect the positions (indexes starting with 1) of source and target and contain additional annotations.

(21) (Sp1) 从 那 时 开 始 这 里 就 成 了 香 港 的 一 个 禁 区 。
(Sp1) Since then , this area has become a prohibited zone in Hong Kong .
19-15(FUN) 17-10(SEM) 7,8-5[DET],6(GIS) 9[COO]-(NTR) 12,13-13,14(SEM)
18-11(SEM) 14[DEP]-12(PDE) 2,5,6-2(FUN) 3,4-3(SEM)
10,11[TEN]-7[TEN],8(GIS)
15,16[MEA]-9(GIF) -4[COO](NTR) -1[MET](MTA) 1[MET]-(MTA)

The separate parts are combined to create a parallel aligned sentence for each line of our corpus. The Chinese segments were segmented into their more common forms typically found in Chinese dictionaries. For instance, '这 里', which has a space between the characters, becomes '这里', with no space ('this area' - in the case of this sentence). This step was performed using the Stanford Chinese Segmenter (Chang et al., 2008);

[28] Some sentences had no alignments and so were unusable and consequently removed.

(Chang et al., 2009c); (Tseng et al., 2005). The word alignments were then adjusted to accommodate the changes.

The final stage of the process involves removing the meta-data and additional annotations in order to match other common word alignment formats. Multiple alignments are split into separate alignment points and then reordered to improve readability. Example 22 is the final version of Example 21 and shows the typical format of the sentences in our corpus.

(22) 从 那时 开始 这里 就 成 了 香港 的 一 个 禁区 。 ||| since then , this area has become a prohibited zone in hong kong . ||| 0-0 1-1 2-0 3-3 3-4 5-5 5-6 6-5 6-6 7-11 7-12 8-10 9-7 10-7 11-8 11-9 12-13

6.2.3 Explicitation Methods

In this section we first outline the process of recovering the implicit information and inserting tokens into our corpus using a heuristic method based on word alignment information. The main goal of such a method is to produce training data for a fully automated method. We then outline our initial fully automated method to predict implicit elements in the data without resorting to word alignment information.

A Heuristic Method to Recover Implicit Elements Here we outline the method of using data from a parallel corpus to identify and target the missing elements. This method relies on word alignments (oracle or automated) to gain knowledge of where the unaligned elements occur in the corpus. This method is suitable for building training corpora, for gaining insight on where implicitation may occur in the data, and for demonstrating the potential impact of making implicit information explicit. However, it cannot be used in practice at decoding time, as translations for the test set (and thus word alignment information) will not be available (they will need to be generated).

To mark implicit elements, we first POS tag[29] the corpus. In this process sentence a) is transformed into sentence b) as shown in Example 23:

[29]For all POS tagging tasks (Chinese and English) we use the Stanford Log-Linear Part-Of-Speech Tagger (Toutanova et al., 2003).

(23) a) 自然 资源 相对 缺乏，

||| natural resources are relatively scarce .

||| 0-0 1-1 2-3 3-4 4-5

b) _自然#NN _资源#NN _相对#AD _缺乏#VV _，#PU

||| natural_JJ resources_NNS are_VBP relatively_RB scarce_JJ ._.

||| 0-0 1-1 2-3 3-4 4-5

The next step retrieves the positions of all the words on the English side that have no corresponding alignment on the Chinese side. In sentence a) the word '**are**' (index position 2) is not aligned to any Chinese counterpart. This can be tagged in one of many ways:

i) are_VBP[30] (both the word and its POS type)

ii) _VBP (a more general POS category token)

iii) are (just the word)

iv) <TOK> (a hold all general place holder token)

Once the element is tagged it is inserted into the Chinese segment. In order to do this, each side of the element is examined to find the nearest aligned English neighbour. The tagged element is then placed, as a token, next to the Chinese counterpart of said neighbour. If both neighbours are equally close, the left neighbour is given preference.

In this case 'resources' (position 1) and 'relatively' (position 3) are aligned to '资源' (resources) and '相对' (relatively), respectively, on the Chinese side. The token is hence inserted immediately after the alignment point for its left neighbour ('资源', resources). Everything to the right of the insertion then moves over. Sentence c) shows the final result after the insertion of the token using the markup in i):

(24) c) 自然 资源 are_VBP 相对 缺乏，||| natural resources are relatively scarce .

A quick check, showed that inserting the word 'are' (with the POS tag removed) into the sentence improves the automated translation[31] (Example 25 is the original, and Example 26 is the modified sentence).

[30] _VBP = Verb, non-3rd person singular present.

[31] Google Translate (https://translate.google.co.uk/)

Table 12: The words and POS elements used in our experiments (Section 6.3).

POS description	*Word coupled with POS tag*	*Counts*
Coordinating conjunctions	and_CC, or_CC, but_CC	13373
Personal pronouns	it_PRP, you_PRP, they_PRP, (s)he_PRP	9672
Subordinating conjunction	if_IN, because_IN	1037
Verb singular present	's_VBZ (3rd per), are_VBP (non-3rd per)	915, 171

(25) 自然资源相对缺乏，

natural resources are relatively scarce . (Human Translation)

Relative lack of natural resources, (MT)

(26) 自然资源 are 相对缺乏，(token replaced with the relevant word)

natural resources are relatively scarce . (Human Translation)

Natural resources are relatively scarce, (MT)[32]

Choosing Insertions Depending on test criteria there are a number of options to consider when making the insertions. Firstly, a choice has to be made as to what POS types to make insertions for (hence the POS tagging step). We can choose to make insertions for every element with no alignment or we can, for example, exclude certain elements, such as punctuation.

For this work we chose to include a specific subset of elements based on the discussion in Section 6.2.1. Table 12 shows a representative list of the POS tagged elements, with their corresponding Penn Treebank descriptions, and POS category frequency counts, that we used in our experiments (Section 6.3). The first three rows relate to our primary focus on DMs and pronouns, whereas the final row includes two elements that are often linked with the pronouns (e.g. 'it's' in Example 27, Section 6.3.3). The final decision on which elements to include was made based upon frequency counts.

With different corpora and languages it may be necessary to experiment with the number and type of tags to include. Some tags may be aligned more often in one language,

[32]This still holds when tested on Google Translate in July 2019. It appears that having the correct '**are**' token enables the model to leverage extra information. An incorrect token breaks the translation.

Table 13: Highlighting the differences between insertions of tokens based on oracle and automated alignments.

Tokens	*Oracle alignments*	*Automated alignments*
and_CC	12350	5015
or_CC	554	95
the_DT	1160	33751

but less often in another. In addition, the method or software used to process the word alignments may also give different results. For our training split of the dataset, using the oracle word alignments, 'and_CC' was inserted 12350 times across 41693 sentences, whilst 'or_CC' was only inserted 554 times. Insertions were made for the POS groups in Table 12 in over 26000 (63%) of the sentences.

Thus far, the focus has been on insertions being made using oracle word alignments. However, we also experimented with automated word alignments, where we created an equivalent corpus using the same insertion rules, but with inserts made based on alignments extracted by Fast-Align (Dyer et al., 2013). Counts for insertions made on the same corpus, but using Fast-Align alignments, vary considerably. For example, 'and_CC' has 5015 insertions (previously 12350) whilst 'or_CC' has 95 (previously 554). Table 13 shows the difference in frequency of insertions made for 'and_CC', 'or_CC', and 'the_DT' using the oracle alignments and automated alignments, respectively.

The word 'the_DT' is included in Table 13 as an additional observation (to be explored in future work) because it does not have a direct equivalent in Chinese. In the GALE corpus 'the' is often merged with the noun it is restricting or modifying. For instance, '城市' ('city') is actually aligned to 'the city'. This method is formally applied to the function words: 'the', 'a', 'an', 'this', 'that' (Li et al., 2009).

Conversely, Fast-Align makes no such distinctions. As a result, when making insertions using the oracle alignments, insertions for 'the_DT' were made 1160 times. When performing insertions on the same data, but using Fast-Align alignments, insertions for 'the_DT' were made on 33751 occasions (roughly 29 times as many). This highlights yet another major difficulty for automatic word alignment tools.

6.2.4 A Method to Predict Implicit Elements

Section 6.2.3 showed that by using heuristics based on word alignments we can locate specific unaligned elements in a sentence. These methods cannot be used for unseen test data at translation time. Our ultimate aim is to use our data with insertions made using this method to train a classifier that predicts whether or not an insertion should occur after a word in a given Chinese sentence, without resorting to any information on the English side.

For our initial model we use CRFsuite (Okazaki, 2007) and our training set of 41693 sentences (annotated automatically with insertions) to build a prediction model, treating the problem as a sequence labelling task. The test set is made up of 1000 sentences (annotated in the same way for evaluation purposes) that do not appear in the training set. Individual sentences are converted into a sequence of tokens (one per line) and each is attached to its POS category and a label signalling whether it precedes an insert or not. As an example the first word from the sentence discussed in Section 6.2.3 would be placed into a file like so: _自然#NN #NN NON .

A template file is then used to describe each word with a number of features. For our initial experiments, the following simple features for each word in the sentence were extracted:

- the preceding two individual words and the following two individual words
- a bigram including the word itself and the word immediately to the left
- a bigram including the word itself and the word immediately to the right
- POS tags for the preceding two words and the following two words
- POS bigrams for the preceding two words through to the following two words
- POS trigrams for the preceding two words (and the word itself) through to the following two words (e.g. POS[-2] | POS[-1] | POS[0] = #DT | #LB | #NN).

The performance of the model is measured using precision and recall. For our test set

that contains 598 insertions, our model labelled 123 (21%) elements as 'PRE' (preceding an insert), with a precision of 84%.

As these are early tests the results are promising, but our future work will need to address two main issues: Firstly, we are currently only predicting 21% of the implicit elements and it is anticipated that experimenting with feature extraction will yield better results. Secondly, we are currently only tagging whether a word precedes an implicit element or not (i.e. no distinction between words).

6.3 Experiments with SMT

In this section we provide experiments using our corpora annotated as per Section 6.2.3 to build SMT systems. The overall aim is to compare SMT systems built and tested with raw parallel data against SMT systems where the source side of the corpus is annotated with place holder information. The corpus annotations are derived from either oracle or automated word alignments. Predicted annotations (Predicted_Inserts) in the source of the test set are produced through a fully automated process, using a classifier trained on oracle alignments. This section also provides a number of examples highlighting how some translations have changed either for better or worse.

6.3.1 Settings and Methodology

Our SMT systems are built using the corpus described in Section 6.2.2. CDEC (Dyer et al., 2010) is used for rule extraction and decoding following the hierarchical phrase-based approach (Chiang, 2007) for Chinese-English translation. We use BLEU (Papineni et al., 2002) as the metric to evaluate the systems. For consistency, default parameters are used during different builds with the only change being the source of the word alignments.

We perform the same experiments twice, with two different splits of the corpus. For each experiment, the corpus is first randomly shuffled. The development and test sets (dev and tst in the table) are then created using the first 2000 sentences (1000 for each) in the shuffled corpus, while the training set is made up of the remaining 41693 sentences. For an oracle build, all sets (dev, tst, and training) include the human created alignments,

whereas for the full automated build, Fast-Alignment (FA) alignments are used. Once the alignment points are added to the sets, each individual sentence has the format shown in Example 22 (Section 6.2.2).

Each experiment consists of five builds:

- **Oracle**: an SMT system built using oracle alignments (no insertions).
- **Baseline_FA**: a baseline SMT system using Fast-Align (FA) (no insertions).
- **Oracle+Inserts**: an oracle SMT system with insertions made using heuristics based on oracle alignments.
- **FA+Inserts**: an automated SMT system with insertions made using heuristics based on Fast-Align alignments.
- **Predicted_Inserts**: an SMT system with insertions made by a classifier using oracle alignments for training the classifier.

Two scores are produced for each of the five builds per experiment, one for the development set and one for the test set.

6.3.2 Results

Table 14 shows BLEU scores for the different experiments with the two different splits of the data. In all cases our experiments have shown that having insertions has a strong positive effect on the scores.

As expected, out of all systems, the Oracle builds perform the best. However, the builds using inserts and Fast-Align (FA+Inserts) show a compelling improvement of up to 1.38 BLEU points over the baseline (Baseline_FA) on the test sets. Similarly, the Oracle builds with inserts (Oracle+Inserts) show a convincing improvement over the plain Oracle builds. More noteworthy is the fact that modest but credible improvements of up to 0.44 are made with our fully automated builds (Predicted_Inserts) over the baseline (Baseline_FA).

Table 14: Examples of the benchmark, baseline, and insertion scores.

(Experiment A)		(Experiment B)	
Build Type	*BLEU*	*Build Type*	*BLEU*
Oracle (dev)	17.81	Oracle (dev)	18.54
Oracle (tst)	**18.34**	Oracle(B) (tst)	18.59
Baseline_FA (dev)	16.59	Baseline_FA (dev)	17.16
Baseline_FA (tst)	**16.76**	Baseline_FA (tst)	**17.02**
Oracle+Inserts (dev)	18.62	Oracle+Inserts (dev)	19.37
Oracle+Inserts (tst)	**19.11**	Oracle+Inserts(tst)	19.38
FA+Inserts (dev)	17.80	FA+Inserts (dev)	18.08
FA+Inserts (tst)	**18.00**	FA+Inserts (tst)	**18.40**
Predicted_Inserts (dev)	16.95	Predicted_Inserts (dev)	17.42
Predicted_Inserts (tst)	**17.20**	Predicted_Inserts (tst)	17.41

6.3.3 Going Beyond BLEU Scores - an Overview of the Output Sentences

BLEU by itself does not give information about what improvements have been made and why, so here we provide some examples taken from our translations, to show the changes. Upon manual inspection of our test translations we noted that a large negative factor was the abundance of out of vocabulary (OOV) words - a possible side effect of the limited sized corpus we used.

Each of the following examples is taken from translations produced by our SMT systems built using inserts based on Fast-Align (FA+Inserts) and has four distinct parts:

i the original sentence (source and target);

ii the source with (automatic) insertions (if any) that appear in the sentence;

iii our (FA+Inserts) system translation (with inserts in the source data);

iv baseline translation (no inserts in the source data).

In each case the ideal output is for item 3 to be a good coherent sentence that closely maps to the target sentence in item 1 and is smoother than the baseline translation shown in item 4.

(27) i 因为 便宜。||| because it 's cheap .

ii 因为it_PRP 's_VBZ 便宜。

iii because **it 's** cheaper .

iv because cheaper .

(28) i 就说 这个 人 长得 像 猴子。||| say, this person looks like a monkey.[33]

ii No direct insertions made in this sentence

iii that is to say **this person** looks like a monkey .

iv that is to say , who looks like a monkey .

(29) i 这 次 会谈 主要 讨论 三 国 在 经贸 文化 等 领域 的 合作，没有 涉及 历史 问题。||| the meeting focused on the three nations ' cooperation in economy , trade and culture , and did not touch on any history problems .

ii 这 次 会谈 主要 讨论 三 国 在 经贸and_CC 文化 等 领域 的 合作，and_CC 没有 涉及 历史 问题。

iii this meeting primarily discuss cooperation in the fields of economics **and** trade, culture, in the three countries , **and** there is no problem involved in history.

iv talks this time will primarily discuss cooperation in areas such as economics and trade , culture , the three countries have on the issue of history .

[33]The reference sentence is quite poor here.

(30) [34]

i 后来 又 说 学生 会 人 太少，没 精力。||| later he said the student association had no energy due to a shortage of hands .

ii 后来he_PRP 又 说 学生 会 人 太少，没 精力。

iii later , **he also** said that the student association people . no , energy .

iv later , people will also said that students 太少 , i did n't energy .

(31) i 中朝 友谊 已经 成为 双方 共同 的 宝贵 财富。||| the friendship between china and north korea has become a precious treasure for the two sides .

ii 中朝and_CC 友谊 已经 成为 双方 共同 的 宝贵 财富。

iii **north korea and** friendship has become the peoples of both sides together .

iv the friendship between china and north korea has become a precious wealth of both sides together .

In Examples 27-30 the sentences translated using data containing inserts are generally much better than the baseline translations. Having inserts appears to affect the overall alignment and decoding process (e.g. weights), so even those sentences without inserts within the actual sentence boundary itself (Example 28) often still show improvements.

Occasionally, having inserts did not help. In Example 31, the baseline translation is clearly better. The insert appears to cause degradation, which could be attributed to conflict with the character pair '中朝' ('zhōng cháo'). By itself '中朝' already has the meaning 'China and North Korea', but the way it is written here is akin to 'sino-DPRK (Democratic People's Republic of Korea)'. That is, the common forms of each country (中国 - China, 北朝鲜 - North Korea) have been truncated and used in a specific (less common) way, which already carries the 'and' information within. Essentially, our insertion of 'and_CC', outside of this tight character pair, introduces extra complexity that is clearly difficult for the MT system to deal with.

[34]The original Chinese sentence in example 8 does not contain the phrase 'shortage of hands' but rather: 人(people) 太少(too few)... An MT system will therefore struggle to produce the actual phrase 'shortage of hands'.

6.3.4 Conclusions

In this section, following a detailed corpus analysis we presented an approach for locating and tagging implicit elements in a parallel aligned corpus. We applied this information to an insertion task that placed proxy tokens for implicit elements into the source data. The data was then used to train SMT systems that were stronger than our baselines.

The source data with the newly inserted elements was also used to train a binary classifier that ultimately was able to predict where implicit elements should occur on unseen data. The data was again used to train SMT systems and the results showed improvements over the baseline.

We faced a barrier with OOV words, which could perhaps be resolved by using a larger dataset. In addition, we observed a strong variance in how items such as function words are treated by oracle and automated alignments. Alignment software lacks the judgement factor of human translation and the gulf in the variance is something that needs to be addressed, or at least, explored.

We believe it would be worth investigating further how much degradation was introduced into the process similar to that shown in Example 31. Our translation system tried to add an additional '**and**' where it wasn't needed. Finding ways to mitigate this kind of damage should lead to even better results.

It is worth noting that we only experimented with a relatively simple set of features. We believe that improving the CRF template and using a wider array of pertinent features will significantly enhance the prediction model, particularly in terms of recall. This, in turn, should lead to further improvements in the quality of translations produced by our SMT systems.

Chapter 7

7 Visualisation Tools

Due to the nature of our work it has been essential to be able to visualise output of the various tools we use in a reliable and consistent manner. This includes viewing, at a granular level, word alignment and metric evaluation output as well as the outputs from different translation systems themselves.

In addition, we had to be able to visualise all the aforesaid output, rapidly with each experiment and in a robust manner where we were not reliant on external services or huge system downloads.

To that end we built, from scratch, two vital tools, which we outline here.

They are:

- WA-Continuum (A word alignment visualisation tool - Section 7.1)
- Vis-Eval (A tool for viewing the output from machine translation systems alongside the relative metric scores - Section 7.2)

7.1 Visualising Word Alignments

WA between pairs of sentences across languages is a key component of many natural language processing tasks. It is commonly used for identifying the translation relationships between words and phrases in parallel sentences from two different languages. Visualising WAs has been a key part of our research throughout and so we developed a tool to

make visualisation of WA much more helpful. Our tool WA-Continuum is designed exclusively for the visualisation of WAs and was initially built to aid research studying WAs and ways to improve them. WA-Continuum relies on the automated mark-up of WAs, as typically produced by other WA tools. Different from most previous work, it presents the alignment information graphically in a WA matrix that can be easily understood by users, as opposed to text connected by lines. The key features of the tool are the ability to visualise WA matrices for multiple parallel aligned sentences simultaneously in a single place (as shown in Chapter 6), coupled with powerful search and selection components to find and inspect particular sentences as required.

7.1.1 Introduction

Automatically generated WA of parallel sentences, as introduced by the IBM models (Brown et al., 1990), is a mapping between source words and target words. It plays a vital role in SMT as the initial step to generate translation rules in most state of the art SMT approaches. It is also widely classed as a valuable linguistic resource for multilingual text processing in general.

Accurate WAs form the basis for constructing probabilistic word or phrase-based translation dictionaries, as well as the generation of more elaborate translation rules, such as hierarchical or syntax-based rules. As WAs improve, it is expected that the translation rules also improve, which, in turn, should lead to better MT.

Our research (as highlighted in Chapter 6) involves a careful study and evaluation of the WA process and aims to develop ways to improve its performance. A substantial part of evaluating WAs often includes human intervention where candidate WAs produced by various software are examined. Consequently, tools to display the alignment information are very important for humans to analyse and readily digest such information.

Various tools have been developed in previous work that enable the visualisation and, in some cases, direct manipulation of WAs. However, none of these tools meet important requirements in our research such as being able to examine WAs for tens and even hundreds of sentences simultaneously in a clear format, and being able to search, shuffle, and

filter those alignments according to desired specific criteria. The WA-Continuum tool was developed to meet this need. It is implemented in Python and outputs to standard HTML files, utilising the powerful properties provided by CSS and JavaScript. As the output file is saved as regular HTML it works with modern web browsers and thus users can make use of many of the features they provide, such as 'search and find'.

The WA-Continuum tool is described in detail throughout as follows: Section 7.1.2 gives a brief overview of existing WA visualisation tools. Section 7.1.3 highlights the technical specification of the WA-Continuum software as well as a number of useful features. Section 7.1.5 presents the conclusion along with a brief overview of future development plans.

7.1.2 Previous Word Alignment Visualisation Tools

With the continuing attention given to SMT and the overarching importance of WAs, various tools have been developed that help evaluate and visualise such alignments by going beyond using text alone. To understand the limitations of text-only visualisation, consider the following Chinese-English example, where the alignments are given in terms of the positions of source-target tokens:

(32) 我 爱 你 ! ||| i love you ! ||| 0-0 1-1 2-2 3-3. (see Figure 5)

For short and simple sentences with numerous 1-to-1 monotone alignments this visualisation style can be sufficient. However, it is certainly not suitable for longer and more complex sentences that may contain more intricate alignments, as per Example 33.

(33) 你好, 我想订两个人的从七月二十七号到三十一号的。||| hello , i 'd like to make a reservation for two in july from the twenty-seventh through the thirty-first . ||| 0-0 1-0 2-1 3-2 4-3 4-4 4-5 5-6 5-7 5-8 5-9 6-10 9-11 11-12 10-13 12-14 12-15 13-15 14-15 15-16 16-17 17-17 15-18 18-19. (see Figure 6)

Figure 5: Showing a simple word alignment for the sentence - I love you!

Using our tool to visualise even this simple sentence already shows how much more intuitive a pictorial representation is. Figure 5 shows the word alignments for the simple sentence given in Example 32. In this case the word order is identical for both sentences in both languages, hence the diagonal line. However, seeing it represented pictorially like this makes the relations very clear.

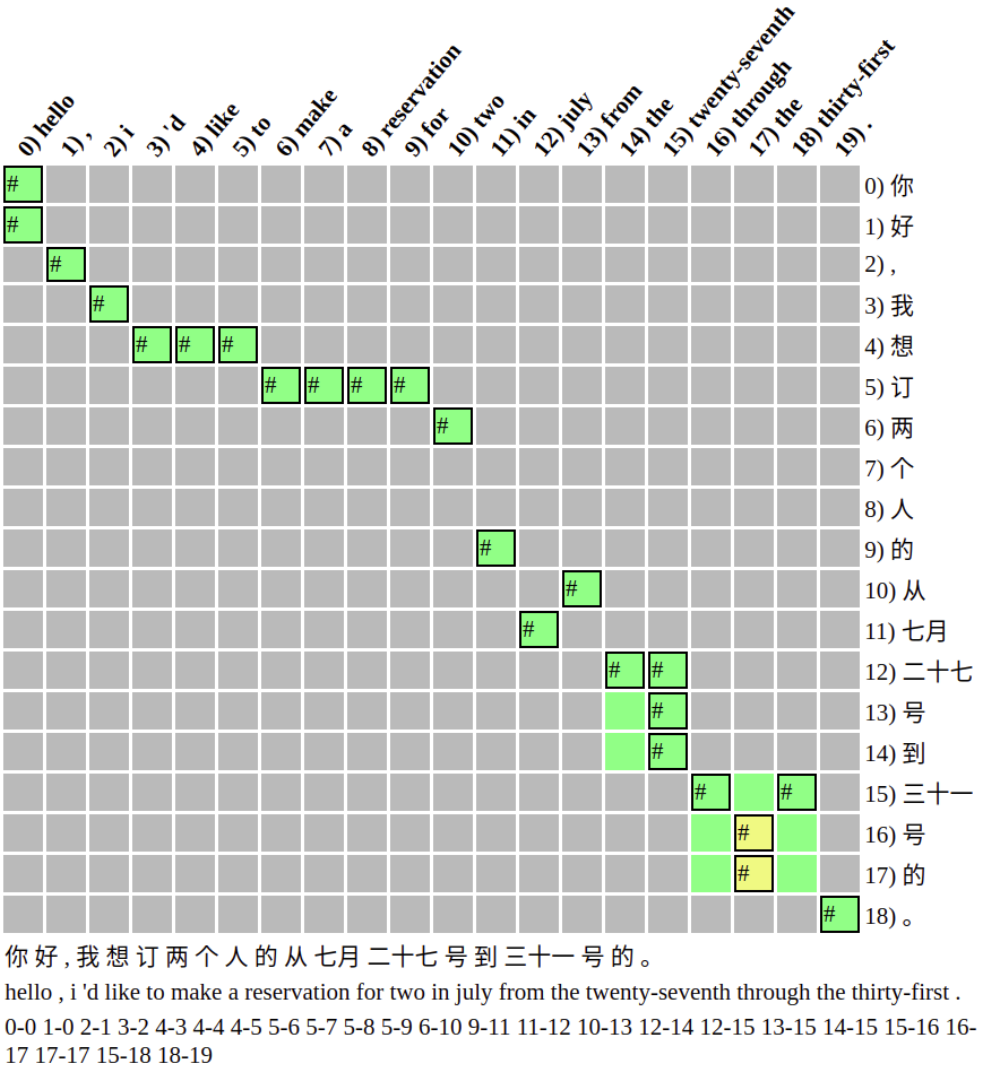

Figure 6: Showing how even complex sentence word alignments become easier to understand.

Previous tools include Cairo (Smith and Jahr, 2000) and VisualLIHLA (Caseli et al., 2008) and have different implementations serving different purposes, but they are usually presented in one of two main visual styles. The earlier styles show alignments by matching words in text boxes, across two sentences, using arrows or lines to make the connections.

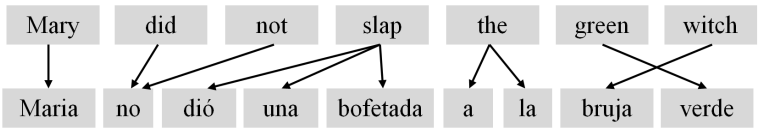

Figure 7: A simple graphical visualisation of WAs for an English-Spanish parallel sentence (Jurafsky and Martin, 2009).

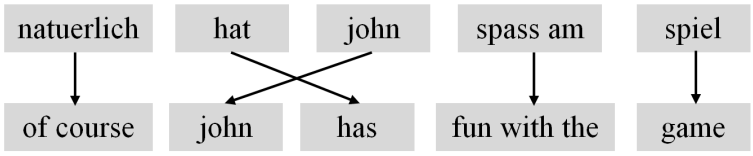

Figure 8: A simple graphical visualisation of WAs for a German-English parallel sentence. The input has been segmented into phrases (Koehn, 2013).

Figure 7 shows an example of this style, where the words in a parallel English and Spanish sentence have been aligned. From the example it can be seen clearly which words map to each other, where reordering occurs (arrows cross), and where phrases are mapped to single words (e.g. 'did not' is mapped to 'no').

Figure 8 shows a similar mapping, but this time it places whole phrases within a single text box and shows both word and phrase alignments. Again, the place where the arrows cross shows some reordering has occurred. The accuracy of the alignments shown in both figures is not a concern, as the tools are purely designed for visualisation purposes. The clarity in how the information is presented, on the other hand, is critical.

The second and perhaps more sophisticated style displays the alignments in a matrix type grid, where the individual columns of the grid map to single elements (words or punctuation marks) in one language and the rows do likewise for single elements in the other language. Figure 9 shows the same parallel sentences as those in Figure 7, but in the grid style. Single blocks show mappings between individual elements (e.g. 'Mary' and 'Maria') whereas multiple blocks appearing in the same row or column tend to show phrases mapping to single words or other phrases (e.g. 'did not' maps to 'no'). As can be seen from Figures 7, 8, and 9 the same information is presented in two different formats, both of which are more intuitive than showing text and word position numbers only.

The tools described so far are static and only show visual representations of WAs.

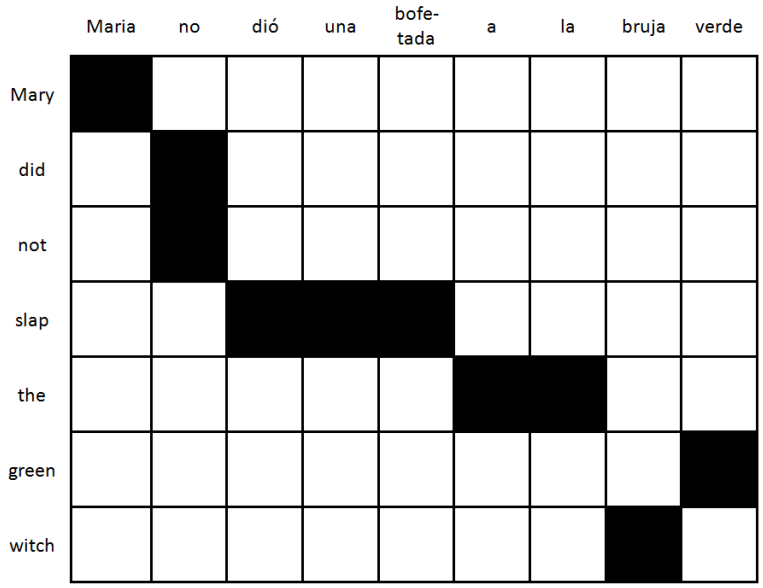

Figure 9: A graphical visualisation of WAs for the given Spanish-English parallel sentence using the matrix format. The columns represent Spanish words whilst the rows represent English words (Jurafsky and Martin, 2009).

Tools such as Yawat (Yet Another Word Alignment Tool) (Germann, 2008) and the SWIFT Aligner (Gilmanov et al., 2014), however, allow the direct manipulation and editing of WAs via graphical interfaces. Picaro – a simple command-line alignment visualisation tool (Riesa, 2011) – uses the grid style to display information. It also has an online demo web page[35] that allows for the demonstration of the tool within a browser for a single parallel sentence. Although Picaro is a relatively simple tool, the visual presentation of the grid format on the demonstration web page is clear and is ideal for quickly understanding WAs. Our research in SMT requires the use of this type of presentation style using the grid format, but with a few more powerful features. Consequently, we had to develop a new tool that had extra features, but maintained the visual appeal and simplicity of the grid format.

7.1.3 Software Features

This section provides an overview of our software including input format and technical specification, as well as a number of the powerful features that we have been using.

[35] http://nlg.isi.edu/demos/picaro/

Input and Technical Specification WA-Continuum is written in Python (version 2.7). The input commands can be typed directly into the command line on Mac, Linux and Windows computers or laptops. They can also be passed as arguments in a number of integrated development environments (IDEs) such as Eclipse[36] or Spyder[37].

The input for the tool should include at least one aligned parallel sentence arranged in the following format:

SOURCE ||| TARGET ||| WAs.

For example:

我 爱 你 ||| i love you ||| 0-0 1-1 2-2

Typically though the input will be a text file containing a list of many such aligned parallel sentences, one per line. The file is read along with an optional user-selected keyword or key-phrase (e.g. -k 'hello' or -k 'as soon as'), which then only returns sentence pairs containing that given word or phrase. Once these commands have been provided, the output is returned as an HTML page, which uses a mixture of HTML, CSS and JavaScript. The page is then automatically opened in the default web browser. This implementation has been successfully tested with a number of modern web browsers including Mozilla Firefox, Internet Explorer 11, Google Chrome, and Opera.

A single web page can show thousands of alignment grids (it has been tested for 10000+ sentences), but despite the fact that the program produces the HTML for the output very quickly, it takes the browser a while to render the page when thousands of grids are involved. We have found through testing that up to 1000 grids can be loaded and rendered fairly quickly (under four seconds on an Intel dual core i3-3220 (3GHZ) computer with 12GB of RAM running Windows 8.1), and so, for performance, we have set the current maximum number of grids to 512 as this is usually enough per search for inspection and evaluation purposes.

A short video showing a demonstration of the WA-Continuum software is available online, see 'Downloading the Tool'.

[36]https://eclipse.org/

[37]https://github.com/spyder-ide/spyder/releases

7.1.4 Features

This section provides an overview of the pertinent features that have been developed and used in our research including: keyword search, phrase search, simple regular expression searches, viewing phrase pairs (minimal bi-phrases), and utilising useful browser features.

For all the given figures in this section exemplifying the WA-continuum software, the individual coordinates for each square in the matrices should be read as row number first, followed by the column number. For Figure 10, the alignment point mapping '因为' to 'because' (as highlighted at the top and right hand side) should be read as alignment point 3-5. The three lines of text below each grid show the source language, target language and WAs as they appear in the input file.

Keyword Searching

As the main aim of the WA-Continuum software is to be able to display clearly WAs for many sentences (possibly the whole corpus), a keyword search was implemented to enable users to select sentences to visualise from the input file, for example, for the analysis of particular constructions such as those using discourse markers.

Figure 10 shows a typical alignment grid returned from using the keyword search 'because'. The '14' in the top left of the figure is an indication that it is the 15th[38] grid for 'because' that appears in the output page. Scrolling up the page will show previous sentences featuring 'because', while scrolling down will show subsequent sentences.

Phrase Searching

This is simply an extension of the keyword search, but by enclosing the search term in quotes it enables the user to input a phrase. For example, a user could easily run the program with the search term 'as soon as' and only results containing that complete phrase

[38]The sentence count starts at 0 to keep it consistent with the alignment point numbering, which also starts at 0.

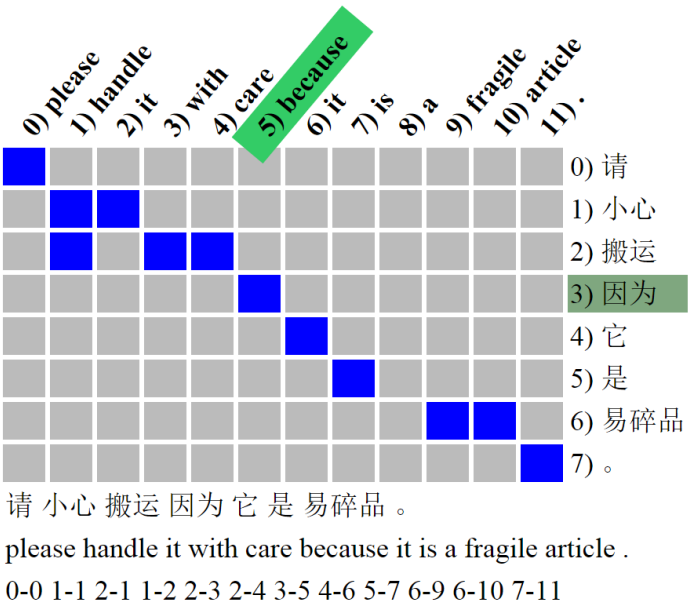

Figure 10: An example of a WA grid returned using the keyword search term 'because'. The cursor was placed over the alignment point for 'because' and '因为' (point 3-5) so the tokens involved in the alignment are highlighted.

will be returned. If the 'as soon as' was typed without the quotes, the tool will return results for the keyword 'as'.

It is worth noting here that the keyword/phrase searches also apply to other alphabets/languages in the input file. For example, a user could do a search using either 'china' (lower case) or '中国'.

Support for Simple Regular Expressions

While keyword and phrase searches are useful tools, if the user is looking for more specific sentences then they can use searches combined with basic regular expressions (RE). Figure 11 is an example of WAs returned using the RE search term 'if.*, then' which is being used to examine sentences containing the if/then conditional. Using the RE search term 'if.*, then' matches any sentence that contains: 'if' followed by any number of characters (.*) followed by a comma and space and finally a 'then'. Being able to use REs makes the search very flexible and helps to pinpoint specific examples.

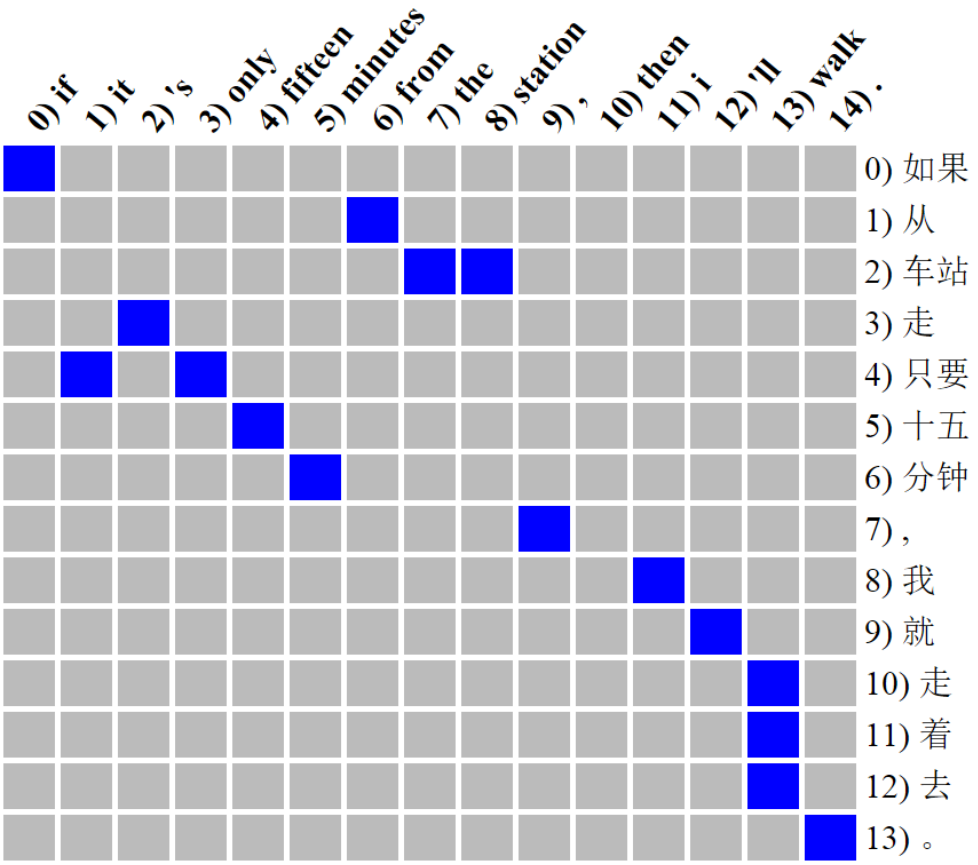

Figure 11: A WA grid returned by using the regular expression search term ‘if.*, then’.

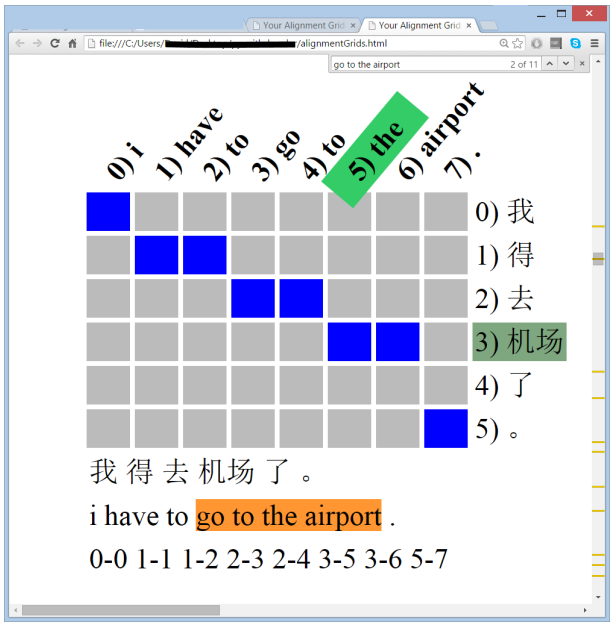

Figure 12: Using the web browser features to search the results. In this case, matches for ‘go to the airport’ are sought.

Using Browser Features

Web browsers often contain many powerful features, but one that is particularly useful for searching the output of tens or hundreds of grids is the ‘search and find’ function. Figure 12 shows a browser search for ‘go to the airport’ being performed on all alignment grids returned by the original command-line keyword search term ‘the’. The figure shows that the sentence being examined is the fifty-fifth one on the page as well as it being the second out of eleven containing matches for ‘go to the airport’. The up and down arrows next to the search term enable the user to jump through the matches on the page with ease. Finally, the small yellow/orange lines on the right hand side show where the other grids containing a match appear on the page.

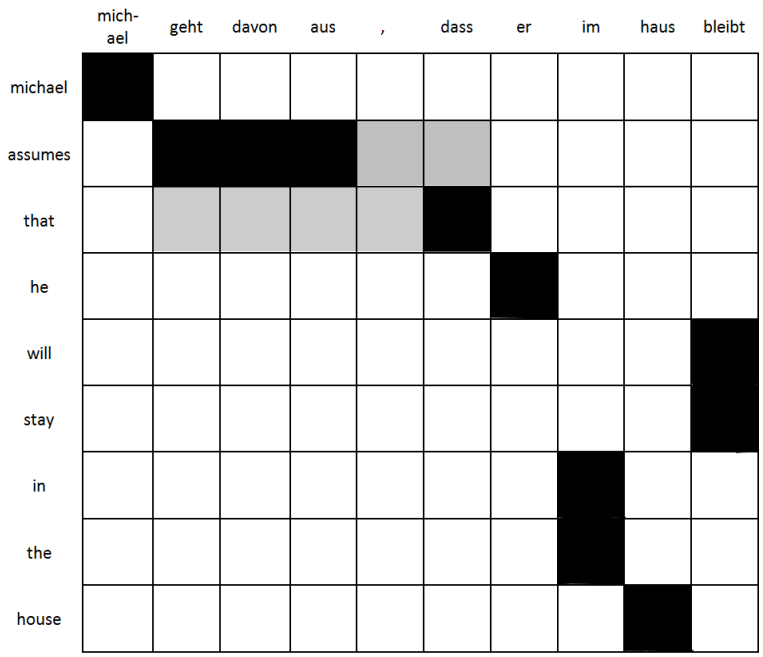

Figure 13: A WA grid showing a phrase pair mapping of ‘assumes that’ to ‘geht davon aus , dass’ Koehn (2013).

Phrase Pairs (Minimal Bi-phrases)

(Koehn, 2013) describes the idea of extracting phrase pairs from word alignments for phrase-based SMT. The reasoning is that if a phrase pair has been identified, it can then be used as evidence for the translation of future occurrences of the phrase. Figure 13 shows an example where ‘assumes that’ has been mapped to ‘geht davon aus , dass’. Using this idea we enabled our software to highlight phrase pairs in order to make it easier to evaluate the WAs, not just for single words, but also for entire phrases. The input file remains the same, but when the optional ‘-b’ switch, for bi-phrases on, is used in the command-line then the tool recursively extracts the phrase pairs at runtime and displays them in the relevant matrices.

Figure 14 shows the first result returned using the phrase search ‘as soon as’ plus the command-line flag ‘-b’, which highlights phrase pairs. Each single block containing the hash symbol represents the actual word alignment points, whereas the large block represents phrase alignments. Phrase alignments will always appear as rectangles and

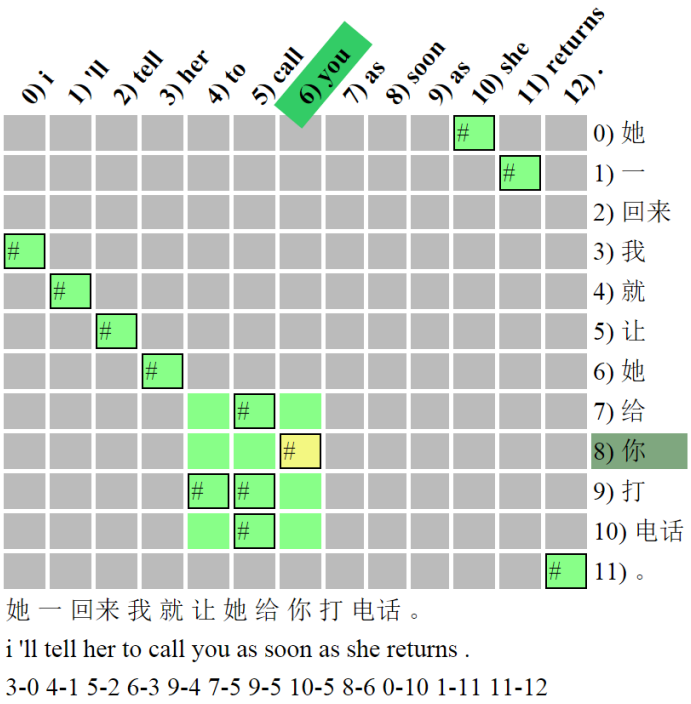

Figure 14: A WA grid showing a phrase pair mapping of ‘to call you’ to ‘给 你 打 电话’.

may include blocks that were not originally aligned (coloured, but no hash symbol). In the context of Figure 14, the English words ‘to call you’ have been mapped as a phrase to ‘给 你 打 电话’ (literally: ‘give you make phone [call]). In this case, quite a good translation.

The process to establish a phrase pair works as follows[39]. If column 5 (‘call’) is examined it is clear that it contains three mappings to rows 7, 9 and 10 respectively. This means that in order to use column 5 in a phrase we must include every alignment point that occurs in the column and by extension those that appear in each of the rows 7, 9 and 10. However, to get from row 7 to row 9 we must also include everything in row 8, and so it goes on in a recursive process.

The phrase ‘to call you’ uses columns 4, 5 and 6. Column 4 has an alignment point at row 9 (9-4). Row 9 in turn also has an alignment point with column 5 (9-5), which then encompasses the other alignment points in column 5 (7-5 and 10-5). As moving through column 5 includes using row 8 then we must also include all alignment points for that row

[39] Also see (Koehn, 2017)

as well, which in this case is in column 6 (8-6). After this, as there are no more alignment points to consider outside of that block, then the phrase is complete. A similar process is applied in Figure 13, which is why the ‘ , ’ in column 4 must be included as part of the phrase ‘geht davon aus , dass’.

Another point worth noting in Figure 14 is that the alignment at point 8-6 (highlighted) mapping ‘你’ to ‘you’ is in a different colour. The reason for this is that the software has been developed to show possible phrases/words that may occur within a larger phrase (nested phrases), as well as being a phrase or single aligned word in its own right. That is, in this case no other item appears in column 6 or row 8 and so the word alignment could be extracted in its own right as a mapping between ‘你’ and ‘you’. None of the other elements that appear in the phrase ‘to call you’ have the same property.

Finally columns 7, 8, 9, and row 2 have no alignment points in them at all. This means that the alignment software has not found suitable alignments for these elements. Using the grid format enables one to spot this issue right away. Based on knowledge of Chinese, we can also quickly spot that the word ‘returns’ (column 11) should be mapped to ‘回来’ (row 2) and ‘as soon as’ (columns 7, 8 and 9) should be mapped to ‘一’ (row 1). These errors would be much harder to spot when examining the alignments in a text only format.

Other Features

A number of other features are available to the user, including the ability to shuffle the results, select a range of matrices, and filter the results to include sentences under a certain length. Extra features such as these are continually being incorporated into the software as the need arises. Furthermore, as the software is open source and well documented, its modular design will enable others to develop and extend the tool with relative ease, and to add further features as required.

Downloading the Tool

WA-Continuum can currently be downloaded from the following location on GitHub: https://github.com/David-Steele/WA-Continuum.

A demonstration video showing how to use many of the described features can also be downloaded from the same repository.

7.1.5 Conclusion

WA-Continuum was designed with one main specific purpose in mind, which is visualising WAs for a large number of sentences at once, making it possible to evaluate them more efficiently. Software enabling the visualisation of WAs has been developed in previous work, and offers a myriad of features including manual editing of WAs and text highlighting. However, none of the tools that we found appeared to offer the full set of functionalities that were required. WA-Continuum builds on the idea of displaying WAs in an intuitive matrix style, whilst making the accessing and searching of large volumes of data a fairly straightforward task.

7.2 Vis-Eval: Visualising Machine Translation System Outputs and their Respective Metric Scores

Machine Translation systems are usually evaluated and compared using automated evaluation metrics such as BLEU and METEOR to score the generated translations against human translations. However, the interaction with the output from the metrics is relatively limited and results are commonly a single score along with a few additional statistics. Whilst this may be enough for system comparison it does not provide much useful feedback or a means for inspecting translations and their respective scores. Vis-Eval Metric Viewer (VEMV) is a tool designed to provide visualisation of multiple evaluation scores so they can be easily interpreted by a user. VEMV takes in the source, reference, and hypothesis files as parameters, and scores the hypotheses using several popular evaluation metrics simultaneously. Scores are produced at both the sentence and dataset level and results are written locally to a series of HTML files that can be viewed on a web browser. The individual scored sentences can easily be inspected using powerful search and selection functions and results can be visualised with graphical representations of the scores and distributions.

7.2.1 Introduction

Automatic evaluation of MT hypotheses is key for system development and comparison. Even though human assessment ultimately provides more reliable and insightful information, automatic evaluation is faster, cheaper, and often considered more consistent.

Many metrics have been proposed for MT that compare system translations against human references, with the most popular being BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), TER (Snover et al., 2006), and, more recently, BEER (Stanojevic and Sima'an, 2014). These and other automatic metrics are often criticised for providing scores that can be non-intuitive and uninformative, especially at the sentence level (Zhang et al., 2004; Song et al., 2013; Bogdan, 2014). Additionally, scores across different metrics can be inconsistent with each other. This inconsistency can be an indicator of linguistic properties of the translations which should be further analysed.

However, multiple metrics are not always used and any discrepancies among them tend to be ignored.

Vis-Eval Metric Viewer (VEMV) was developed as a tool bearing in mind the aforementioned issues. It enables rapid evaluation of MT output, currently employing up to eight popular metrics. Results can be easily inspected (using a typical web browser) especially at the segment level, with each sentence (source, reference, and hypothesis) clearly presented in interactive score tables, along with informative statistical graphs. No server or internet connection is required. Only readily available packages or libraries are used locally.

Ultimately VEMV is an accessible utility that can be run quickly and easily on all the main platforms.

Before describing the technical specification of the VEMV tool and its features in Section 7.2.3, we give an overview of existing metric visualisation tools in Section 7.2.2.

7.2.2 Related Tools

Several tools have been developed to visualise the output of MT evaluation metrics that go beyond displaying just single scores and/or a few statistics.

Despite its criticisms and limitations, BLEU is still regarded as the *de facto* evaluation metric used for rating and comparing MT systems. It was one of the earliest metrics to assert a high enough correlation with human judgments.

Interactive BLEU (iBleu) (Madnani, 2011) is a visual and interactive scoring environment that uses BLEU. Users select the source, reference, and hypothesis files using a Graphical User Interface (GUI) and these are scored. The dataset BLEU score is shown alongside a bar chart of sentence scores. Users can select one of the sentences by clicking on the individual bars in the chart. When a sentence is selected its source and hypothesis translation is also shown, along with the standard BLEU statistics (e.g. score and n-gram information for the segment). Whilst iBLEU does provide some interactivity, using the graph itself to choose the sentences is not very intuitive. In addition the tool provides results for only one metric.

METEOR is another popular metric used to compute sentence and dataset-level scores based on reference and hypothesis files. One of its main components is to word-align the words in the reference and hypothesis. The Meteor-X-Ray tool (Denkowski, 2014) (Denkowski and Lavie, 2014) generates graphical output with visualisation of word alignments and scores. The alignments and score distributions are used to generate simple graphs (output to PDF). Whilst the graphs do provide extra information there is little in the way of interactivity.

MT-ComparEval (Klejch et al., 2015) is a different evaluation visualisation tool, available to be used online[40] or downloaded locally. Its primary function is to enable users, via a GUI, to compare two (or more) MT system outputs, using BLEU as the evaluation metric. It shows results at both the sentence and dataset level highlighting confirmed, improving, and worsening n-grams for each MT system with respect to the other. Sentence-level metrics (also n-gram) include precision, recall, and F-Measure information as well as score differences between MT systems for a given sentence. Users can upload their own datasets to view sentence-level and dataset scores, albeit with a very limited choice of metrics. The GUI provides some interaction with the evaluation results and users can make a number of preference selections via check boxes.

The Asiya Toolkit (Giménez and Màrquez, 2010) is a visualisation tool that can be used online or as a stand-alone tool. It offers a comprehensive suite of metrics, including many linguistically motivated ones. Unless the goal is to run a large number of metrics, the download version is not very practical. It relies on many external tools such as syntactic and semantic parsers. The online tool[41] aims to offer a more practical solution, where users can upload their translations. The tool offers a module for sentence-level inspection through interactive tables. Some basic dataset-level graphs are also displayed and can be used to compare system scores.

In comparison to the other software described here, VEMV is a light yet powerful utility, which offers a wide range of metrics and can be easily extended to add other metrics. It has a very specific purpose in that it is designed for rapid and simple use

[40] http://wmt.ufal.cz/

[41] At the time of writing the online version did not work.

locally, without the need for servers, access to the internet, uploads, or large installs. Users can quickly get evaluation scores from a number of mainstream metrics and view them immediately in easily navigable interactive score tables. We contend that currently there is no other similar tool that is lightweight and offers this functionality and simplicity.

7.2.3 Vis-Eval Metric Viewer Software & Features

This section provides an overview of the VEMV software and outlines the required input parameters, technical specifications, and highlights a number of the useful features.

The Software VEMV is essentially a multi-metric evaluation tool that uses three tokenised text files (source, reference, and hypothesis) as input parameters and scores the hypothesis translation (MT system output) using up to eight popular metrics: BLEU, MT-Eval[42] (MT NIST & MT BLEU), METEOR, BEER, TER, Word Error Rate (WER), and Edit Distance (E-Dist).[43] All results are displayed via easily navigable web pages that include details of all sentences and scores (shown in interactive score tables - Figure 15). A number of graphs showing various score distributions are also created.

The key aims of VEMV are to make the evaluation of MT system translations easy to undertake and to provide a wide range of feedback that helps the user to inspect how well their system performed, both at the sentence and dataset level.

☑ POS ☑ REF Length ☑ Sentence ☑ Sen Bleu ☑ MT Bleu ☑ MT Nist ☑ METEOR ☑ BEER ☑ TER ☑ WER ☑ Edit Dist

POS	REF Length	Sentence	Sen Bleu	MT Bleu	MT NIST	MET-EOR	BEER	TER	WER (Score)	E Dist (Score)
1	9 (44)	SRC: 你 有 那 个 症状 多 长 时间 了 ? REF: how long have you been having that symptom ? HYP: how long have you been that symptom ?	0.6102	0.6102	11.1177 (0.7713)	0.4613	0.818	0.1111	1 (0.8889)	7 (0.8409)
2	11 (41)	SRC: 我 在 哪 坐 去 波士顿 的 巴士 ? REF: where can i catch a bus to go to boston ? HYP: where do i get the bus for boston ?	0.1886	0.096	3.6523 (0.2529)	0.3052	0.4682	0.5455	6 (0.4545)	16 (0.6098)

Figure 15: A screenshot of an interactive score table showing two example sentences and their respective scores.

[42] https://www.nist.gov/

[43] A WER like metric that calculates the Levenshtein (edit) distance between two strings, but at the character level.

7.2.4 Input and Technical Specification

VEMV is written in Python 3 (also compatible with Python 2.7). To run the tool, the following software needs to be installed:

- Python >= 2.7 (required)
- NLTK[44] >= 3.2.4 (required)
- Numpy (required)
- Matplotlib / Seaborn (optional - for graphs)
- Perl (optional - for MT BLEU, MT NIST)
- Java (optional - for METEOR, BEER, TER)

With the minimum required items installed the software will generate scores for standard BLEU, WER, and E-Dist. The optional items enable a user to run a wider range of metrics and produce nearly 200 graphs during evaluation.

The input commands to run the software can be typed directly into the command line on any platform, or passed as arguments in an interactive development environment (IDE) such as Spyder.[45]

Once the software has been run (see Section 7.2.7), a folder containing all of the generated HTML, text, and image files is produced. A user will typically explore the output by opening the 'main.html' file in a browser (Chrome, Firefox, and Opera have been tested) and navigating it like with any (offline) website. The text files contain the output for the various metric scores and can be inspected in detail. The graphs are output as image files (PNGs), which are primarily viewed in the HTML pages, but can also be used separately for reports (e.g. Figure 17 in Section 7.2.6)

[44] http://www.nltk.org
[45] https://github.com/spyder-ide/spyder/releases

7.2.5 Main Features

Here we outline some key features of the Vis-Eval Metric Viewer tool:

Scoring with Multiple Evaluation Metrics

Currently VEMV uses eight evaluation metrics to score individual sentences and the whole document. All results are shown side by side for comparison purposes and can be inspected at a granular level (Figure 15).

A glance at the two sentences in Figure 15 already provides numerous points for analysis. For example, the MT in sentence 2 is a long way from the reference and receives low metric scores. However, whilst not identical to the reference, the MT is correct and could be interchanged with the reference without losing meaning. For sentence 1 the MT is only a single word away from the reference and receives good scores, (much higher than sentence 2) although the meaning is incorrect. The interactive display enables the user to easily examine such phenomena in a given dataset.

Clear and Easily Navigable Output

The main output is shown as a series of web pages and can be viewed in modern browsers. The browsers themselves also have a number of powerful built-in functions, such as page search, which are applicable to any of the output pages, adding an extra layer of functionality.

The output consists of easily navigable interactive score tables and graphs, logically organised across web pages. The tool includes its own search facility (for target and source sentences) and the option to show or hide metric scores to aid clarity, especially useful for comparing only a selection of metrics. All of the segment level metric scores can be sorted according to the metric of interest.

Results Saved Locally

Once scored, the generated text files, images, and HTML pages are saved locally in a number of organised folders. The majority of the text files are made up from the standard raw output of the metrics themselves. The image files are statistical graphs produced from the metric scores. Both the text and image files can be inspected directly on a metric by metric basis and used for reference. The VEMV tool brings together the text and images in the HTML files to form the main viewable output.

Runtime User Options

The minimal default settings will quickly produce scores for standard BLEU, WER and E-Dist. Numerous parameters can be set on the command line enabling the user to choose any or all of the additional metrics and whether or not to generate graphs.

A number of the metrics (especially BLEU and METEOR) have a plethora of parameters, which can be selected. To avoid the need for complex command line inputs the metric level parameters can be placed in an easily editable text-based configuration file, which in turn is passed to the command line.

In addition, the user can choose which metric will be the dominant one for sorting and display purposes (the default is BLEU) and there is an option for selecting how many score bins or pages to use to show the sentences. The default is 100 pages (one for every percentage point), but some users may prefer fewer pages (e.g. 10 or 20) in order to simplify the main interface and general navigation.

An accessibility flag has also been added. It removes some of the colour formatting from the displays making it easier for users with visual impairments (e.g colour blindness).

7.2.6 Viewing the Actual Output

Figure 16 shows the main page of the software. In this case all eight metrics were used as shown by the mini graph icons. Each of these mini graph icons acts as a link. Ten score

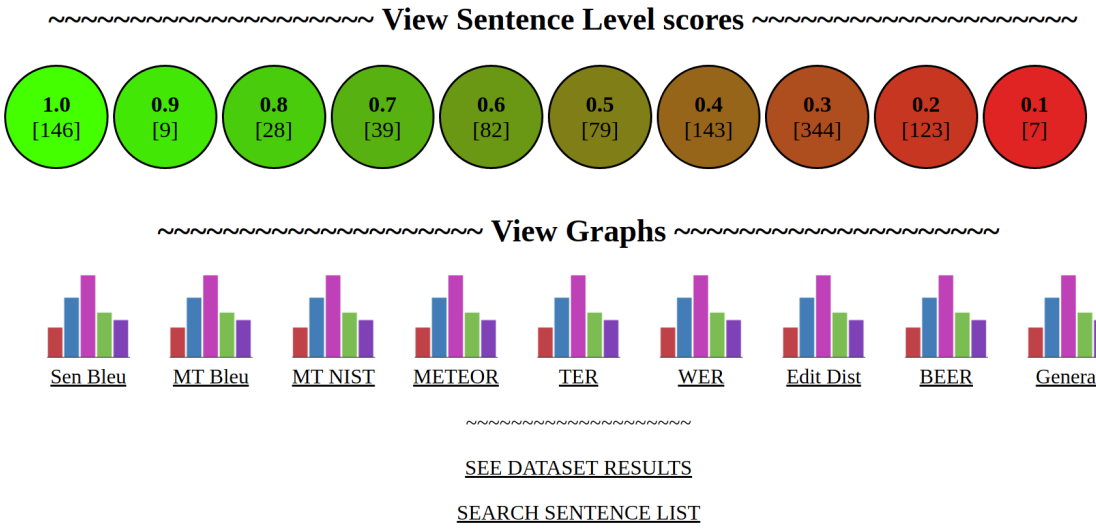

Figure 16: A screenshot of the VisEval Metric Viewer main page.

bins (circular icons) were selected as a parameter.

Users can click on any of the links/icons to navigate to the various pages. Clicking on the circular icons opens the sentence level score pages (Figure 15) showing individual sentences with a given score. Clicking on the mini graph icons takes the user to the graph display web pages for the respective metrics or the general document-wide statistics. Figure 17, for example, is a metric graph showing the distribution of standard BLEU scores for the dataset. In this case the chart in Figure 17 would be accessed by clicking on the very left hand mini graph icon on the main page shown in Figure 16.

7.2.7 Downloading and Running the Tool

Vis-Eval Metric Viewer can currently be downloaded from the following location on GitHub: https://github.com/David-Steele/VisEval_Metric_Viewer.

The associated README file provides instructions on how to get started with using the tool, and what to do if you run into any problems.

In terms of hardware requirements, a computer with at least 2GB of RAM and 300MB of available storage is needed to run the software.

A short video demonstration of these and other features of the Vis-Eval Metric Viewer software can be found online at: https://youtu.be/nUmdlXGYeMs.

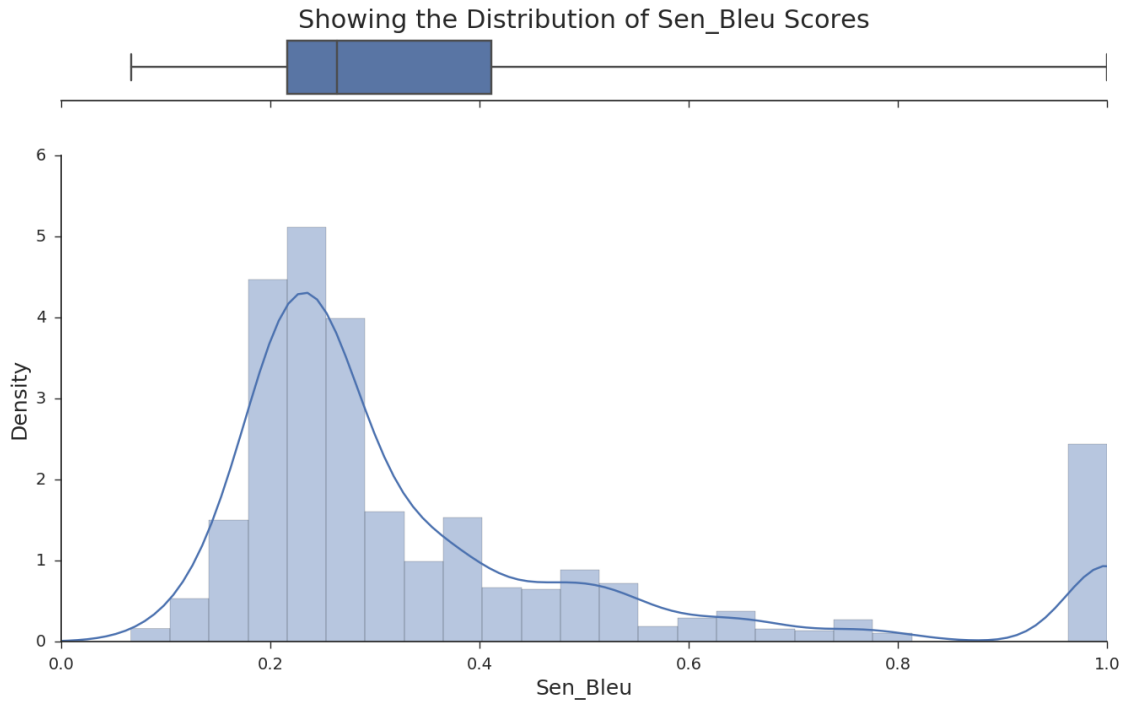

Figure 17: A graph showing the distribution of standard BLEU scores.

7.2.8 Conclusion

The Vis-Eval Metric Viewer tool was designed with three main aims:

- To provide a useful tool that is easy to install (using readily available packages), and simple to use and run on a local machine without the need for a server or internet connection.
- To offer a single place for scoring translations using multiple popular metrics.
- To provide in depth visual feedback making it easy to examine segment level metric scores.

The tool offers a light weight solution that makes it easy to compare multiple-metric scores in a clear manner. Feedback can be interactively explored and searched rapidly with ease, whilst numerous graphs provide additional information. The tool can be run locally on any platform. All results are saved in self-contained folders for easy access.

This tool is ideal for use with any MT paradigm as it operates on the output files. In this research it has been used successfully for showing both SMT and NMT output. We believe it therefore has longevity as it can be used in the increasingly adopted world of NMT.

Section 8.3 highlights some of the additions we should like to make to the software, to provide even more useful functionality.

Chapter 8

8 Conclusions

In this final chapter we summarise the work that we have covered throughout the thesis and evaluate to what extent we have achieved our aims as set out in Chapter 1. Finally, we present some possible directions for future work.

8.1 Summary

For many years SMT was the leading MT paradigm, but, at the time of writing, it was being replaced by NMT, which has now been widely adopted. However, SMT is still commonly used and there have been vast developments in SMT models in recent years.

Sections 2.1 and 2.2 in Chapter 2 presented the chain of events that led to SMT being so strong for so long. This is our justification for choosing to focus in depth on and around this topic for our research.

To put this thesis into context there has been increasing effort and resources going towards addressing discourse in MT, in particular cohesion, our primary area of interest. Sections 4, 4.2, and 4.3 in Chapter 4 showed the nature and type of effort applied, and presented some of the advances that have made SMT so strong. Chapter 2 also presented a brief history of NLP culminating in reasons for Chinese being such an important language for translation, and hence being the subject of many targeted projects. We also note that Chinese and English are considered to be a very difficult language pair, making an interesting choice for our research.

With regard to cohesion in SMT our contribution has been focussed on discourse connectives, discourse markers, and discourse relations on a sentential level.

Our initial goal was to establish the important information, cues, and links that discourse connectives and, to some extent, other devices hold or signify within text. In Chapter 3 we presented an overview of the function of cohesion and cohesive devices within language and discourse, and highlighted specific elements that are very different when considering Chinese and English (our language pair of choice). These differences in usage of cohesive devices enabled us to find numerous challenging sentences (as per Chapters 1 and 5) that strong SMT models (and even NMT models) struggle to translate. However, it was important to delve much deeper. Therefore, the importance of cohesion, established in Chapter 3 guided us into a focussed analysis of specific discourse elements in Chinese.

In Chapter 5 we used important and well established bilingual corpora to produce a robust corpus analysis highlighting the usage and spread of select discourse devices. The outcome of this analysis gave insight into the divergence in the usage of discourse markers between Chinese and English, and enabled us to appreciate the problems they could cause for SMT systems (e.g. omission and long range connections). We also discussed how explicitly marking connectives within Chinese sentences helped to improve translations of the given examples.

In Chapter 6 we built on this idea and, through a more in depth corpus analysis, started to explore concrete methods of (heuristically) inserting special placeholder tokens (e.g. <TOK>) into the source language of the given bilingual corpora. The 'adjusted' corpora could then be used in the formal word alignment process in order to improve the alignment rules, which in turn led to much smoother final translations in a number of cases.

However, the real power of this approach could only be realised if this became an automated process and could be used without the target language being available. Hence, we built a prediction model (Section 6.2) trained on gold standard word aligned corpora that could predict where placeholder tokens for select connectives could be placed in our test text. As above, the word alignment process would then leverage this extra information

in order to improve the alignment rules. The results of using this methodology showed strong improvements over the baseline, ultimately leading to better translations.

A significant part of our research involved being able to visualise the results of our methods. With so much focus being placed on improving word alignments it was imperative that we could quickly manually inspect hundreds of word alignments both before and after our changes to observe the improvements (or damage). In Chapter 7, Section 7.1 we presented WA-Continuum, our word alignment visualisation tool, and detailed the reasons that it was so important and useful for our research.

Following on from this we also had to have a way of exploring, in detail, our translation system outputs so we could quickly visually compare actual final translations both from before and after applying our methods. In addition, we wanted to explore various metric scores so we could observe the changes and explore the overall effects across the translation text. In Chapter 7, Section 7.2 we presented our Vis-Eval software that was created for this very purposes and demonstrated its effectiveness. Using this tool we could quickly explore any sentence from possibly thousands and look at its scores given by a number of competing metrics. We could then use the tool to explore various language phenomena and the effect of our changes, enabling us to reach conclusions as to how effective our methodology actually was. As an added beneficial side effect we actually stumbled upon other interesting language phenomena and observed patterns of weaknesses across the evaluation metrics.

8.2 Evaluation of Aims

In Chapter 1, Section 1.3 we presented two main research questions (listed again here for reference):

1. What are the limitations of current strong SMT approaches (e.g. hierarchical tree-based) in terms of handling cohesive devices?

2. How can we better model cohesive devices within sentences?

Across the course of this thesis we believe we presented a comprehensive analysis of the limitations SMT systems have when dealing with cohesion, and in particular, cohesive devices. We presented an in depth review of related work (Chapter 4) looking at challenges that cohesive devices present to SMT, and then we focussed on our language pair of choice, using well established corpora to investigate potential problems and test SMT models in order to explore observed weaknesses. In Chapters 5 and 6 we built upon these findings and explored the limitation at a granular level, using heuristics to make changes in order to further explore weaknesses.

With regard to better modelling of cohesive devices we extended this work to produce a prediction model that could deal with problematic cases and be used as part of an overall SMT pipeline to improve final translations. In addition, we were also able to evidence our improvements through voluminous examples and robust results from our myriad of experiments.

Therefore, we respectfully submit that we achieved our aims from initial preliminary general analysis through to a more specific in depth survey resulting in a final piece that better modelled cohesive devices in SMT.

However, our main disappointment with our approach is in the timing and how our investigation into SMT came at a point when there was the start of a strong move toward using NMT.

8.3 Future Work

Here, we outline a number of avenues for future work to develop our findings in this thesis.

For SMT We presented a prediction model built using a gold standard corpus and focussing on a small number of discourse markers. Two obvious ways to progress this work are:

1. To include a wider range of corpora, from the vast number available, to be investigated and processed in order to leverage even more discourse information and enhance the model.

2. To extend the number of discourse connectives used in the model in order to capture a much greater variety of discourse relations and further enhance translations.

For NMT One of the primary ways we should like to progress our work is by applying our methodology to NMT. We may need to alter some of the steps and move away from word alignments, but the main idea would largely be the same. That is, use processes that capture discourse relations and somehow incorporate the connections into the neural network architecture.

Whilst Example 7 showed that the problems of discourse relations in MT are partially solved, Examples 8 and 9 demonstrated cases where NMT still struggles with discourse relations and connectives on some level. Clearly there is scope to explore this further and examine new ways to leverage cohesive information into the process.

Visualisation Software In this thesis we presented two substantial pieces of software that we developed from scratch. Whilst there is always room for development of both items we certainly see a strong future for our Vis-Eval tool as it can be (and has been) applied to the evaluation of results for any MT system. It works on the format of the translation output regardless of the architecture that generated it and so is equally effective for both SMT and NMT alike.

Important changes we should like to make are:

1. To incorporate a variety of extra evaluation metrics (over and above the eight we already have).

2. To add interactivity to the graphs (e.g. zoom functions) to make them dynamic.

3. To use the saved final outputs of the software in a combined way so we can actually compare results and evaluations from two (or more) MT systems.

One final point of note is that whilst using the Vis-Eval tool to examine our results we also identified numerous weaknesses and differences between how individual metrics scored the various sentences. It would be interesting to explore this further and possibly create a new metric based on any findings. Or in the very least, develop ways to strengthen existing metrics.

9 Appendix A - Publications

- Steele, D. and Specia, L. (2014), Divergences in the Usage of Discourse Markers in English and Mandarin Chinese. In Text, Speech and Dialogue (17th International Conference TSD), pages 189–200, Brno, Czech Republic. TSD

- Steele, D. and Specia, L. (2015), WA-Continuum: Visualisng Word Alignments across Multiple Parallel Sentences Simultaneously. In ACL-IJCNLP, Beijing, China. Association for Computational Linguistics

- Steele, D (2015), Improving the Translation of Discourse Markers for Chinese into English, In Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 110 – 117, Denver, Colorado, Association for Computational Linguistics,

- Steele, D. and Specia, L. (2016), Predicting and Using Implicit Discourse Elements in Chinese-English Translation, In Proceedings of the 19th Annual Conference of the European Association for Machine Translation, pages 305–317, EAMT

- Steele, D and Specia, L. (2018), Vis-Eval Metric Viewer: A Visualisation Tool for Inspecting and Evaluating Metric Scores of Machine Translation Output. In Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations), New Orleans, LA. NA-ACL

- Steele, D., Sim Smith, K., and Specia, L. (2015), Sheffield Systems for the Finnish-English WMT Translation Task. In Tenth Workshop on Statistical Machine Translation, pages 172–176, Lisbon, Portugal. Association for Computational Linguistics

References

A. Bies, M. Ferguson, K. Katz, R. MacIntyre, V. Tredinnick, G.M. Kim, M.A. Marcinkiewicz, and B. Schasberger. 2002. Bracketing guidelines for treebank ii style. In *Penn Treebank Project*.

B. Bogdan. 2014. Automated mt evaluation metrics and their limitations. *Tradumàtica: tecnologies de la traducció*, 1(12):464–470.

P. Brown, J. Cocke, S. Della, V. Della, P. Jelinek, J. Lafferty, R. Mercer, and P Roosin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):76–85.

P. Brown, S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. 1993a. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19:263–311.

P.F. Brown, V.J. Della Pietra, S.A. Della Pietra, and R.L. Mercer. 1993b. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

B. Cartoni, S. Zufferey, and Meyer T. 2013. Annotating the meaning of discourse connectives by looking at their translation: The translation-spotting technique. *Dialogue and Discourse : Beyond semantics: the challenges of annotating pragmatic and discourse phenomena*, 4(2):65–86.

H. Caseli, F. Gomes, T. Pardo, and M. Nunes. 2008. Visuallihla: The visual online tool for lexical alignment. In *XIV Brazilian Symposium on Multimedia and the Web*, pages 378–380, Vila Velha, Brazil.

M. Cettolo, C. Girardi, and M. Federico. 2012. Wit3: Web inventory of transcribed and translated talks. In *European Association for Machine Translation*, pages 261–268, Trento, Italy.

P. Chang, D. Jurafsky, and C. Manning. 2009a. Disambiguating "de" for chinese-english machine translation. In *4th Workshop on Statistical Machine Translation, pages 215-223, Athens, Greece.*

P. Chang, H. Tseng, D. Jurafsky, and C.D Manning. 2009b. Discriminative reordering with chinese grammatical relations features. In *Third Workshop on Syntax and Structure in Statistical Translation.*

P.C. Chang, M. Gally, and C. Manning. 2008. Optimizing chinese word segmentation for machine translation performance. In *Association for Computational Linguistics 2008 Third Workshop on Statistical Machine Translation*. Association for Computational Linguistics.

P.C. Chang, H. Tseng, D. Jurafsky, and C.D. Manning. 2009c. Discriminative reordering with chinese grammatical relations features. In *Third Workshop on Syntax and Structure in Statistical Translation*.

D. Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

F.Y Chuang. 2017. Discourse markers - how are paragraphs linked together? Online; Accessed: 15/11/2018.

T. Chung and D. Gildea. 2010. Effects of empty categories on machine translation. In *Conference on Empirical Methods in Natural Language Processing*, pages 636–645.

A. Clark, C. Fox, and S. Lappin. 2013. *The Handbook of Computational Linguistics and Natural Language Processing.* Wiley-Blackwell, Chichester, UK.

L. Danlos and C. Roze. 2011. Traduction (automatique) des connecteurs de discours. In *Traitement Automatique des Langues Naturelles*, Montpellier, France.

M. Denkowski. 2014. Meteor 1.5: Automatic machine translation evaluation system. Online; Accessed on 16/07/2019.

M. Denkowski and A. Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

C. Dyer, V. Chahuneau, and Smith N.A. 2013. A simple, fast and effective reparameterization of ibm model 2. In *Conference of the North American Chapter of the Association for Computational Linguistics*, Atlanta, US.

C. Dyer, A. Lopez, J. Ganitkevitch, J. Weese, F. Ture, P. Blunsom, H. Setiawan, V. Eidelman, and P. Resnik. 2010. Cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Association for Computational Linguistics*.

J.R. Finkel, T. Grenager, and C. Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *43nd Annual Meeting of the Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

K. Forbes, E. Miltsakaki, R. Prasad, A. Sarkar, A. Joshi, and B. Webber. 2001. D-ltag system -discourse parsing with a lexicalised tree adjoining grammar. In *ESSLI 2001 Workshop on Information Structure, Discourse Structure, and Discourse Semantics*, Helsinki, Finland.

W.N Francis and H. Cucera. 1964. *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers.* Brown University, Providence, Rhode Island.

Q. Gao. 2008. Word order in mandarin: Reading and speaking. In *20th North American Conference on Chinese Linguistics (NACCL-20)*, Ohio, USA.

U. Germann. 2008. Yawat: Yet another word alignment tool. In *ACL-08: HLT Demo Session*, pages 20–23, Columbus, Ohio. Association for Computational Linguistics.

T. Gilmanov, O. Scrivner, and S. Kubler. 2014. Swift aligner, a multifunctional tool

for parallel corpora: Visualization, word alignment, and (morpho)-syntactic cross-language transfer. In *Language Resources and Evaluation Conference*, pages 2913–2919, Reykjavik, Iceland.

J. Giménez and L. Màrquez. 2010. Asiya: An open toolkit for automatic machine translation (meta-)evaluation. *The Prague Bulletin of Mathematical Linguistics*, 94:77–86.

N. Hajlaoui and A. Popescu-Belis. 2013. Translating english discourse connectives into arabic: a corpus-based analysis and an evaluation metric. In *CAASL4 Workshop at AMTA (Fourth Workshop on Computational Approaches to Arabic Script-based Languages*, pages 1–8, San Diego, CA.

M. A. K. Halliday and R. Hasan. 1994. *Cohesion in English.* Longman, London, UK.

M.A.K Halliday and R. Hasan. 1976. *Cohesion in English (English Language Series*. Longmen, London, London, UK.

C. Hardmeier. 2012. Discourse in statistical machine translation: A survey and a case study. In *Discours - Revue de linguistique, psycholinguistique et informatique*, Caen. Presses Universitaires de Caen.

C. Hardmeier. 2014. *Discourse in Statistical Machine Translation.* Uppsala University, Elanders Sverige, Sweden.

C. Hardmeier, S. Stymne, J. Tiedemann, and J. Nivre. 2013. Docent: A document-level decoder for phrase-based statistical machine translation. In *51st Annual Meeting of the Association for Computational Linguistics.*, pages 193–198, Sofia, Bulgaria.

U. Hermjakob, K. Knight, and H Daumé III. 2008. Name translation in statistical machine translation: Learning when to transliterate. In *ACL-08: HLT*, pages 389–397, Columbus, Ohio, USA.

J. Huang. 1989. Pro-drop in chinese a generalized control approach. In *Jaeggli, O and Safir, K. (editors): The NUll Subject Parameter*, pages 185–214. Kluwer Academic Publishers.

J. Hutchins. 2005. The history of machine translation in a nutshell. Online; Accessed 13/05/2014.

B. Hutchinson. 2004. Acquiring the meaning of discourse markers. In *42nd meeting of Association for Computational Linguistics, Main Volume*, pages 684 – 691, Barcelona, Spain.

D. Jurafsky and J. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Rcognition: 2nd Edition.* Pearson Prentice Hall, Upper Saddle River, New Jersey.

O. Klejch, E. Avramidis, A. Burchardt, and M. Popel. 2015. MT-compareval: Graphical evaluation interface for machine translation development. *The Prague Bulletin of Mathematical Linguistics*, 1(104):63–74.

P. Koehn. 2013. *Statistical Machine Translation.* Cambridge University Press, Cambridge UK.

P. Koehn. 2017. Moses: Smt system, user manual and code guide. Online; Accessed on 16/07/2019.

A. Kroch and A. Taylor. 2000. The penn-helsinki parsed corpus of middle english, second edition. Online; Accessed on 22/07/2019.

H. Kucera and F Winthrop-Nelson. 1967. *Computational Analysis of Present Day American English.* Brown University Press, US.

R. Levy and C.D Manning. 2003. Is it harder to parse chinese, or the chinese treebank? In *Association for Computational Linguistics*, pages 439 – 446.

H. Li and B. Yuan. 1998. Chinese word segmentation. In *Language, Information and Computation (PACLIC12*, pages 212–217.

X. Li, N. Ge, and S. Strassel. 2009. *Tagging Guidelines for Chinese-English Word Alignment - Version 1.0, Linguistic Data Consortium.* Online; Accessed 19/07/19.

Y. Li. 2008. Sensitive positions and chinese complex sentences: A comparative perspective. *Journal of Chinese Language and Computing*, 18(2):47–59.

A. Lopez. 2008. Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40(3). Article No.8.

A. Lopez and M. Post. 2013. Beyond bitext: Five open problems in machine translation. In *Workshop on Twenty Years of Bitext: Association for Computational Linguistics*, Seattle, Washington, USA.

X. Luo and B. Zhao. 2011. A statistical tree annotator and its applications. In *49th annual meeting Association for Computational Linguistics*, pages 1230–1238, Portland, Organ.

Macmillan. 2009. Macmillan dictionary. Online; accessed 2014.

N. Madnani. 2011. ibleu: Interactively debugging & scoring statistical machine translation systems. In *Proceedings of the Fifth IEEE International Conference on Semantic Computing*, pages 213–214.

M.H. Manser, editor. 2009. *Oxford Chinese Dictionary: English-Chinese Chinese-English.* Oxford University Press, UK.

D. Marcu and W. Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing. (EMNLP).*

T. Meyer and A. Popescu-Belis. 2012. Using sense-labelled discourse connectives for statistical machine translation. In *EACL Joint Workshop on Exploiting Synergies between IR and MT, and Hybrid Approaches to MT (ESIRMTHyTra)*, pages 129–138, Avignon, France.

T. Meyer and B. Webber. 2013. Implicitation of discourse connectives in (machine) translation. In *Workshop on Discourse in Machine Translation (DiscoMT)*, pages 19–26, Sofia, Bulgaria.

R. Mitkov. 2004. *The Oxford Handbook of Computational Linguistics.* Oxford University Press, New York.

T.P. Nguyen, A. Shimazu, T.B Ho, M.L. Nguyen, and V.V. Nguyen. 2008. A tree-to-string phrase-based model for statistical machine translation. In *CoNLL 2008: In: 12th Conference on Computational Natural Language Learning*, pages 14–150, Manchester.

R. Nordquist. 2014. About education - collocation (words). Online; Accessed 12/09/2014.

E. Nyberg, T. Mitamura, and J. Carbonell. 1997. The kant machine translation system: From research and development to initial deployment.

F.J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

N. Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs). Online; Accessed 2015.

J. Olive, C. Christianson, and J. McCary. 2011. *Handbook of Natural Language Processing and Machine Translation.* Springer, New York.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *40th ACL*, pages 311–318, Philadelphia, PA.

M. Paul. 2009. Overview of the iwslt 2009 evaluation campaign. In *IWSLT*.

E. Pitler and A. Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *ACL-IJCNLP 2009 (47th Annual Meeting of the ACL and 4th International Joint Conference on NLP of the AFNLP), Short Papers*, pages 13–16, Singapore.

Y. Po-Ching and D. Rimmington. 1998. *Chinese: Intermediate Chinese, A Grammar and Workbook.* Routledge, London, UK.

Y. Po-Ching and D. Rimmington. 2004. *A Comprehensive Grammar.* Routledge, London, UK.

Y. Po-Ching and D. Rimmington. 2010. *Chinese: An Essential Grammar (2nd Edition).* Routledge, London, UK.

A. Popescu-Belis, B. Cartoni, A. Gesmundo, J. Henderson, C. Hulea, P. Merlo, T. Meyer, J. Moeschler, and S. Zufferey. 2011. Improving mt coherence through text-level processing of input texts: the comtis project. In *Tralogy 2011 (Translation Careers and Technologies: Convergence Points for the Future)*, Paris, France.

A. Popescu-Belis, T. Meyer, J. Liyanapathirana, B. Cartoni, and S. Zufferey. 2012. Discourse-level annotation over europarl for machine translation: Connectives and pronouns. In *Language Resources and Evaluation Conference*, pages 2716–2720, Istanbul, Turkey.

W.J. Rapaport. 2005. The turing test: Verbal behavior as the hallmark of intelligence. *Computational Linguistics*, 31(3).

J. Riesa. 2011. Picaro: A simple command-line alignment visualisation tool. Online; Accessed 2015.

Roget's Thesauraus. 2014. Thesauraus.com. roget's 21st century thesaurus, third edition. Online; accessed 2014.

C. Ross. 2011. *Chinese – Demystified.* McGraw Hill, London, UK.

C. Ross and J. Sheng Ma. 2006. *Modern Mandarin Chinese Grammar.* Routledge, London, UK.

L. Russo. 2011. Étude inter-langues de la distribution et des ambiguïtés syntaxiques des pronoms. In *M. Lafourcade and V. Prince (eds.), TALN 2011/RECITAL 2011 (18e conférence annuelle sur le Traitement automatique des langues naturelles*, pages 279–284.

A. Saluja, C Dyer, and S.B. Cohen. 2014. Latent-variable synchronous cfgs for hierarchical translation. In *Empirical methods in Natural language processing (EMNLP)*, pages 1953–1964, Doha, Qatar.

C.E Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379 – 423.

A. Singla and N. Agrawal. 2009. *Using Named Entity Recognition to Improve Machine Translation.* Online; Accessed 20/07/2019.

N.A. Smith and M.E Jahr. 2000. Cairo: An alignment visualisation tool. In *Language Resources and Evaluation Conference*, Athens, Greece.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

X. Song, T. Cohn, and L. Specia. 2013. Bleu deconstructed: Designing a better mt evaluation metric. *International Journal of Computational Linguistics and Applications*, 4(2):29–44.

L. Specia, M.G.V. Nunes, and M. Stevenson. 2005. Exploiting rules for word sense disambiguation in machine translation. *Procesamiento del Lenguaje Natural*, 35:171–178.

Lucia Specia. 2013. Statistical machine translation. In Sivaji Bandyopadhyay, Sudip Kumar Naskar, and Asif Ekbal, editors, *Emerging Applications of Natural Language Processing: Concepts and New Research*, chapter 4, pages 74–109. IGI Global.

M. Stanojevic and K. Sima'an. 2014. Beer: Better evaluation as ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages "414–419". "Association for Computational Linguistics".

D. Steele. 2015. Improving the translation of discourse markers for chinese into english. In *NAACL-HLT (ACL) 2015 Student Research Workshop (SRW)*, pages 311–318, Denver, Colorado.

D. Steele and L. Specia. 2014. Divergences in the usage of discourse markers in english and mandarin chinese. In *(TSD) Lecture Notes in Computer Science*, pages 189–200, Berlin Heidelberg. Springer.

D. Steele and L. Specia. 2015. Wa-continuum: Visualising word alignments across multiple parallel sentences simultaneously. In *ACL-IJCNLP 2015 System Demonstrations*, Association for Computational Linguistics, pages 121–126, Beijing, China.

D. Steele and L. Specia. 2016. Predicting and using implicit discourse elements in chinese-english translation. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 305–317.

D. Steele and L. Specia. 2018. Vis-eval metric viewer: A visualisation tool for inspecting and evaluating metric scores of machine translation output. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 71–75, New Orleans, LA.

M. Swan and B. Smith. 2004. *Learner English (2nd Edition)*. Cambridge University Press, Cambridge, UK.

Systran. 2014. Systran : What is machine translation? rule-based machine translation technology [available online]. Online; Accessed on 20/04/2014.

T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto. 2002. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Language Resources and Evaluation Conference*, pages 147–152, Las Palmas, Spain.

A. Taylor, M. Marcus, and B. Santorini. 2003. The penn treebank: An overview. *Abeillé A. (eds) Treebanks. Text, Speech and Language Technology*, 20.

Chinese Teachers. 2010. The conjunction 2010. Online; Accessed 2014.

K. Toutanova, D. Klein, C. Manning, and Y. Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL-HLT*, pages 252–259.

H. Tseng, Chang P.C., G. Andrew, D. Jurafsky, and C. Manning. 2005. A conditional random field word segmenter. In *Fourth SIGHAN Workshop on Chinese Language Processing*.

B. Tsou, W. Gao, T. Lai, and S. Chan. 1999. Applying machine learning to identify chinese discourse markers. In *International Conference on Information, Intelligence and Systems*, Chania Crete, Greece.

M. Tu, Y. Zhou, and C. Zong. 2014. Enhancing grammatical cohesion: Generating transitional expressions for smt. In *52nd annual meeting of the Association for Computational Linguistics*, Baltimore, USA.

P. Tung and D. Pollard. 1994. *Character Text For Colloquial Chinese: Simplified Character Version.* University of London.

J. Véronis and P. Langlais. 2000. Evaluation of parallel text alignment systems: The arcade project. In *Parallel Text Processing*, pages 369–388. Kluwer Academic Publishers, Text Speech and Language Technology Series.

C. Wang and L. Huang. 2006. Grammaticalisation of connectives in mandarin chinese: A corpus-based study. *Language and Linguistics*, 7(4):991–1016.

W. Weaver. 1949. Memorandum on translation. Online; accessed 2019 (in the Rockerfeller Foudation Archives).

J. Weizenbaum. 1966. Eliza - a computer program for the study of natural language communication between man and machine. *Communications for the Association for Computing Machinery*, 9(1):36–45.

B. Wong and C. Kit. 2012. Extending machine translation evaluation metrics with lexical cohesion to document level. In *2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1060–1068, Jeju Island, Korea.

D. Wu. 2005. Statistical machine translation part ii: Tree-based smt. In *International Joint Conference on Natural Language Processing*.

J. Wu. 2014. Shifts of cohesive devices in english-chinese translation. *Theory and practice in Language Studies*, 4(8):1659–1664.

Y. Wu, M. Schuster, Z. Chen, Q.V. Le, M. Norouzi, M. Macherey, M. Krikun, Y. Cao, Q. Gao, and K. Macherey. 2016. *Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation.* Google (preprint).

B. Xiang, X. Luo, and B. Zho. 2013. Enlisting the ghost: Modeling empty categories for machine translation. In *51st annual meeting of Association for Computational Linguistics*, pages 822–831, Bulgaria.

T. Xiao, J. Zhu, S. Yao, and H. Zhang. 2011. Document-level consistency verification in machine translation. In *2011 MT summit XIII*, pages 131–138, Xiamen, China.

D. Xiong, B. Guosheng, Z. Min, L. Yajuan, and L. Qun. 2013a. Modelling lexical cohesion for document-level machine translation. In *Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI-13)*, Beijing, China. 2013(a).

D. Xiong, Ding Y., Z. Min, and Chew Lim Tan. 2013b. Lexical chain based cohesion models for document-level statistical machine translation. In *2013 Conference on Empirical Methods in Natural Language Processing*, pages 1563–1573. 2013(b).

J. Xu and R. Bock. 2011. Combination of alternative word segmentations for chinese machine translation. *DARPA Global Autonomous Language Exploitation*.

N. Xue. 2005. Annotating discourse connectives in the chinese treebank. In *Association for Computational Linguistics Workshop on Frontiers in Corpus Annotation 2: Pie in the Sky.*

N. Xue and F. Xia. 2000. The bracketing guidelines for the penn chinese treebank 3.0.

Y. Yang and N. Xue. 2010. Chasing the ghost: Recovering empty categories in the chinese treebank. In *23rd International Conference on Computational Linguistics*, pages 1382–1390, Beijing, China.

F. Yung. 2014. Towards a discourse relation-aware approach for chinese-english machine translation. In *Association for Computational Linguistics, Student Research Workshop*, pages 18–25, Baltimore, Maryland USA.

Y. Zhang, S. Vogel, and A. Waibel. 2004. Interpreting bleu/nist scores: How much improvement do we need to have a better system. In *Proceedings of Language Resources and Evaluation*, pages 2051–2054.

Z. Zhang, T. Cohn, and F. Ciravegna. 2013. Topic-orientated words as features for named entity recognition. In *Computational Linguistics and Intelligent Text Processing. Lecture Notes in Computer Science*, volume 7816, pages 304–316.

S. Zuffery and L. Degand. 2013. Annotating the meaning of discourse connectives in multilingual corpora. *Corpus Linguistics and Linguistic Theory.*, 0(0):1–24.