

An Investigation into Automatic Translation  
of Prepositions in IT Technical  
Documentation from English to Chinese

**Yanli Sun**

A dissertation submitted to Dublin City University in fulfilment  
of the requirements for the degree of

Doctor of Philosophy

School of Applied Language and Intercultural Studies

Supervisors: Dr. Sharon O'Brien; Dr. Minako O'Hagan;

Dr. Fred Hollowood

November 2010

I hereby certify that this material, which I now submit for assessment on the program of study leading to the award of Doctor of Philosophy is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: \_\_\_\_\_ (Candidate) ID No.: \_\_\_\_\_ Date: \_\_\_\_\_

## **Acknowledgements**

The completion of this project has been blessed with the help of many people. First of all, I would love to express my deep gratitude and thanks to my supervisors, Dr. Sharon O'Brien, Dr. Minako O'Hagan from DCU and Dr. Fred Hollowood from Symantec (Ireland) for sharing their ideas with me and giving me time and comments. They always encouraged me to explore new areas and gave me guidance and suggestions with utmost patience. And not only that, their encouragement also helped me to be a more confident and positive person in life. I am also indebted to them for their information on local culture. Especially to Fred, thank you for your jokes and lessons on driving, food and wine, etc.

A special thanks to Dr. Johann Roturier and Dr. Nora Aranberri for putting up with my ignorant questions over the years. Thank you for being there to help and encourage me. It has been an extremely delightful honour to work with you.

My thanks goes to the Innovation Partnerships Programme of Enterprise Ireland and to Symantec Ireland for their generous funding of the project. I would like to acknowledge the colleagues and friends both at Symantec and at DCU for their invaluable company. Thanks to all the warm and friendly Irish people for making my stay here happy and enjoyable.

I would also love to thank my parents for their support through this project. Thanks to my dearest sister and lovely niece for the laughter they gave me and for taking care of the parents while I was away. Finally and most importantly, I would love to thank my husband Dr. Yanjun Ma, who is also my best friend and the best listener in the world. Thank you for making my life meaningful and complete.

## Contents

Acknowledgements .....	iii
List of Tables .....	vii
List of Figures .....	x
List of Abbreviations .....	xii
Abstract.....	xiv
Chapter 1: Introduction.....	1
1.1 Research Background .....	1
1.2 Research Questions.....	4
1.3 Structure of the Thesis .....	5
Chapter 2: Machine Translation and MT Evaluation.....	7
2.1 Machine Translation .....	7
2.1.1 Rule-based Machine Translation .....	8
2.1.2 Data-driven Machine Translation .....	14
2.1.3 Hybrid Machine Translation.....	23
2.2 Evaluation of Machine Translation .....	25
2.2.1 Human Evaluation.....	26
2.2.2 Automatic Evaluation.....	32
2.2.3 Meta-evaluation of Human and Automatic Evaluation.....	40
2.2.4 MT Evaluation Campaigns .....	43
2.3 Summary .....	45
Chapter 3: Contextualising the Research Question - Translation of Prepositions .....	47
3.1 Errors in MT Output .....	48
3.2 English and Chinese Prepositions.....	53
3.2.1 English Prepositions .....	53
3.2.2 Chinese Prepositions .....	59
3.3 Setting up a Preposition Error Typology .....	62
3.4 Research Questions.....	66
3.5 Further Context.....	67
3.6 Summary .....	71
Chapter 4: Methodology.....	73
4.1 MT Systems Used.....	74
4.2 Evaluation – Measurement Validity and Reliability .....	75
4.3 Corpus Design .....	78
4.3.1 Major Principles.....	78

4.3.2	Preposition Corpus and Sample Extraction .....	81
4.4	Other Research Design Issues .....	86
4.5	Pilot Tests.....	87
4.5.1	Testing of Domain-Specific UDs.....	88
4.5.2	Testing of CL .....	90
4.5.3	Testing of Automatic S&R .....	91
4.5.4	Testing of SPE .....	93
4.6	Summary .....	94
Chapter 5: Error Analysis of the Translation of Prepositions.....		96
5.1	Experiment Set-up .....	96
5.2	Evaluation Results .....	99
5.2.1	Question 1: Are all prepositions translated incorrectly? .....	100
5.2.2	Question 2: How frequent is each error? .....	103
5.2.3	Question 3: What are the errors pertaining to each preposition? 105	
5.3	Summary .....	106
Chapter 6: Statistical Post-editing .....		108
6.1	Experiment Set-up .....	109
6.1.1	Building and Modifying an SPE Module .....	110
6.1.2	Evaluation Preparation .....	117
6.2	Evaluation Results .....	120
6.2.1	Reliability of Human Evaluation .....	120
6.2.2	Results - Translation of Prepositions .....	123
6.2.3	Results - Translation of Sentences .....	131
6.2.4	Further Exploration .....	138
6.3	Summary .....	143
Chapter 7: Dictionary Customisation and Source Pre-Processing.....		145
7.1	Automated Preposition Dictionary Extraction .....	146
7.1.1	Experiment Set-up.....	147
7.1.2	Translation Evaluation .....	150
7.2	Statistical Source Reconstruction .....	157
7.2.1	Rationale.....	159
7.2.2	Experiment Set-up.....	163
7.2.3	Results .....	167
7.3	Summary .....	176
Chapter 8: Comparison between RBMT, SMT and SPE .....		178
8.1	Experiment Set-up .....	179
8.1.1	MT Systems .....	179
8.1.2	Corpora.....	181
8.1.3	Obtaining Translations .....	184
8.2	Preparing Human Evaluation .....	185
8.3	Results.....	191
8.3.1	Inter- and Intra-evaluator Correlation .....	192
8.3.2	System Level Comparison.....	194
8.3.3	Preposition Level Comparison.....	198

8.3.4	Linguistic Analysis.....	199
8.4	Summary .....	205
Chapter 9: Conclusion .....		207
9.1	Important Findings .....	208
9.2	Limitations and Future Research.....	212
9.3	Closing Words.....	214
References.....		215
Appendices.....		231
Appendix A – Questionnaire for the First Human Evaluation .....		232
Appendix B – Instructions for the First Human Evaluation .....		233
Appendix C – Instructions for the Second Human Evaluation.....		238
Appendix D – Questionnaire for the Second Human Evaluation .....		241
Appendix E – Instructions for the Third Human Evaluation.....		242
Appendix F – Questionnaire for the Third Human Evaluation.....		245

## List of Tables

Table 2.1: Divergences of Chinese segmentation.....	18
Table 2.2: Pros and cons of SMT, RBMT and EBMT.....	23
Table 2.3: Interpretation of fluency and accuracy scores.....	29
Table 3.1: Summary of errors reported by internal translators.....	50
Table 4.1: Frequency of each preposition in the corpus.....	83
Table 4.2: The top ten frequent prepositions and their relative frequencies .....	83
Table 4.3: Frequencies of the top ten frequent prepositions in other corpora .....	84
Table 4.4: Distribution of entries in the Symantec User Dictionary.....	89
Table 4.5: GTM scores and number of errors in the translations .....	93
Table 5.1: Number of prepositions needing PE vs. those not needing PE .....	101
Table 5.2: Percentages of prepositions being mistranslated.....	102
Table 5.3: Number of each error assigned by the evaluators .....	103
Table 5.4: Inter-evaluator agreement for each error type.....	104
Table 5.5: Distribution of errors in the translation of each preposition .....	105
Table 6.1: Preliminary training, tuning and test corpora.....	110
Table 6.2: Notations for corpora used for SPE.....	110
Table 6.3: Notation for the monolingual phrase table of the SPE module.....	111
Table 6.4: Notations for Baseline and SPED.....	112
Table 6.5: Notations for bilingual and preposition phrase table.....	113
Table 6.6: Notations for SPEP and SPEF.....	116
Table 6.7: Agreement level and inter-evaluator correlation (sentence) .....	121
Table 6.8: Agreement level and inter-evaluator correlation (preposition).....	121
Table 6.9: Number and percent of times that each system was top-ranked (preposition level).....	123
Table 6.10: The percent of times that a system was judged to be better than any other system (preposition level) .....	124

Table 6.11: Pair wise significance test (preposition level).....	125
Table 6.12: Number and percentage of shared translations by any two systems (preposition level).....	125
Table 6.13: Improvement and degradation of each system compared to the Baseline.....	127
Table 6.14: Automatic evaluation scores of Baseline and SPEF/P/D.....	131
Table 6.15: Number and percent of times that each system was top-ranked (sentence level).....	132
Table 6.16: The percent of times that a system is judged as better than any other system (sentence level) .....	133
Table 6.17: Kappa correlation between automatic and human evaluation at sentence level .....	136
Table 6.18: Number of ties (and percentages) assigned by each evaluator.....	136
Table 6.19: Consistency level between automatic score and human ranking ....	137
Table 6.20: Percent of ties within each score difference interval.....	140
Table 6.21: Percent of times that one system was evaluated as better than the others in the post-hoc test .....	142
Table 6.22: Consistency level between automatic score and human evaluation in the post-hoc test.....	143
Table 7.1: Number of entries of each preposition extracted from the phrase table .....	149
Table 7.2: Automatic evaluation scores of the translations with/without the supplementary preposition dictionary .....	151
Table 7.3: Pilot samples for the comparison of back- and forward-translation..	160
Table 7.4: Automatic evaluation scores of the back- and forward-translations .	161
Table 7.5: In-domain training corpus and test set.....	165
Table 7.6: Three mixed-domain training corpora .....	166
Table 7.7: Automatic evaluation scores of pre-processed translations .....	167
Table 7.8: Number and examples of insertions, deletions and substitutions.....	171
Table 7.9: Distribution of the pre-processed sentences within three categories.	174
Table 8.1: Fuzzy matching between training and test corpora .....	182



Table 8.2: Training and tuning corpora for SMT.....	183
Table 8.3: Training and tuning corpora for SPE.....	183
Table 8.4: Number of pairs of translations for human evaluation .....	190
Table 8.5: Inter-evaluator correlation.....	192
Table 8.6: Intra-evaluator correlation.....	193
Table 8.7: Automatic evaluation scores for the four systems.....	194
Table 8.8: Percent of times that the column system is judged as better than the row system (sentence level) .....	195
Table 8.9: Correlation between human evaluation and GTM .....	196
Table 8.10: Refined correlation between human and GTM .....	197
Table 8.11: Percent of times that the column system was judged as better than the row system (preposition level) .....	198
Table 8.12: Frequency of the two categories at sentence level .....	199
Table 8.13: Frequency of the two categories at preposition level .....	202

## List of Figures

Figure 2.1: Flowchart of an RBMT system.....	10
Figure 2.2: An entry from a Moses phrase table.....	20
Figure 2.3: Word alignment information from SL to TL (0 represents the first word).....	20
Figure 2.4: Word alignment information from TL to SL (0 represents the first word).....	21
Figure 2.5: (GTM) Bitext grid between an MT output and the reference.....	35
Figure 3.1: Error typology for an RBMT system .....	48
Figure 3.2: Error typology of an SMT system.....	49
Figure 3.3: Five levels of syntactic structures (Koehn 2003: 2).....	52
Figure 3.4: Systran's (v.6) instruction sample for encoding dictionary entries....	68
Figure 4.1: A sample rule defined in acrolinx IQ .....	81
Figure 4.2: A screenshot of the output file of acrolinx IQ .....	82
Figure 4.3: Distribution of entries in the Symantec User Dictionary .....	89
Figure 5.1: Sample sentence in the evaluation sheet (error analysis-task 1).....	98
Figure 5.2: Sample sentence in the evaluation sheet (error analysis-task 2).....	98
Figure 5.3: Excel Kappa calculator template.....	100
Figure 5.4: Error distribution among the 447 prepositions .....	104
Figure 6.1: Flowchart of the first two steps in the process of modifying SPE...	112
Figure 6.2: Flowchart of steps 3 to 5 in the process of modifying SPE.....	116
Figure 6.3: Screenshot of preposition level evaluation.....	117
Figure 6.4: Screenshot of sentence level evaluation .....	118
Figure 6.5: Sample of the human rankings and automatic evaluation scores.....	134
Figure 6.6: Sample of scores to ranks transformation.....	135
Figure 6.7: Percentage of ties assigned within each score interval.....	139
Figure 7.1: Pipeline for extracting prepositional entries .....	147

Figure 7.2: Automated extracted preposition dictionary.....	150
Figure 7.3: Improvements, degradations and equal translations.....	156
Figure 7.4: Baseline translation from the original source sentence .....	169
Figure 7.5: $MT_{mixed}^{large}$ from pre-processed source sentence.....	169
Figure 8.1: Sample of the human evaluation sheet .....	190

## List of Abbreviations

ALPAC	Automatic Language Processing Advisory Committee
APE	Automatic Post-editing
BLEU	Bilingual Evaluation Understudy
BNC	British National Corpus
CL	Controlled Language
COBUILD	Collins Birmingham University International Language Database
DARPA	The Defence Advanced Research Projects Agency
EAGLES	Expert Advisory Group on Language Engineering Standards
EBMT	Example-Based Machine Translation
FEMTI	Framework for the Evaluation of Machine Translation in ISLE
GTM	General Text Matcher
IPCITI	International Postgraduate Conference in Translation and Interpreting
ISLE	International Standards for Language Engineering
IT	Information Technology
JEIDA	The Japan Electronic Industry Development Association
LDC	Linguistic Data Consortium
LOB	Lancaster-Oslo-Bergen
LREC	International Conference on Language Resources and Evaluation
LSP	Language for Special Purpose
MT	Machine Translation
NIST	The National Institute of Standards and Technology
NP	Noun Phrase
PB-SMT	Phrase-Based Statistical Machine Translation
PE	Post-editing
PKU	Peking University
POS	Part of Speech
PP	Prepositional Phrase
RBMT	Rule-Based Machine Translation

RE	Regular Expressions
S&R	Search and Replace
SL	Source Language
SMT	Statistical Machine Translation
SPE	Statistical Post-editing
TAUS	Translation Automation User Society
TER	Translation Edit Rate
TL	Target Language
TM	Translation Memory
UD	User Dictionary
WMT	Workshop on Machine Translation

## **Abstract**

Machine Translation (MT) technology has been widely used in the localisation industry to boost the productivity of professional translators. However, given the high quality of translation expected, the translation performance of an MT system in isolation is often less than satisfactory due to various errors. This study focuses on translation of prepositions from English into Chinese within technical documents in an industrial localisation context. The aim of the study is to reveal the salient errors in the translation of prepositions and to explore possible methods to remedy these errors.

This study first examines which prepositions were handled unsatisfactorily by the MT system in the study (Systran version 6). Based on this information, three new approaches are proposed to improve the translation of prepositions. All approaches attempt to make use of the strengths of the two most popular MT architectures at the moment: Rule-Based MT (RBMT) and Statistical MT (SMT). The approaches include: first, building an automatic preposition dictionary for the RBMT system; second, exploring and modifying the process of Statistical Post-Editing (SPE) and third, pre-processing the source texts to better suit the RBMT system. Overall evaluation results (either human evaluation or automatic evaluation or both) show the potential of our new approaches in improving the translation of prepositions. The current study also compares some of the state-of-the-art MT systems to reveal which translation architecture should be preferred, especially in regard to achieving better translation of prepositions. Finally, one of the important outcomes of the research is the proposal of a new function of automatic metrics in assisting researchers to obtain more valid or purpose-specific human evaluation results.

# Chapter 1: Introduction

Automatically translating from one natural language into another language using computer technologies (which is known as Machine Translation (MT)) has a long history. Various translation systems have been proposed to date. These systems can now be divided into two broad categories, rule-based or data-driven systems. As the name suggests, a rule-based system uses manually crafted linguistic rules to control the translation process. Data-driven systems require large data sets where translation patterns are learnt automatically by a system. It has been claimed that MT systems are useful for many purposes if the users wish to glean general information (Hutchins 2002). However, raw MT translation does not usually meet the required standards in an industrial context where the translated text is intended to be a final product ready for the customers' scrutiny. While MT has continued to improve over the years, some issues remain unresolved. The current work is rooted in an industrial context, and aims to tackle one of the well-known challenges faced by an RBMT system – translation of prepositions.

## 1.1 Research Background

The funding for this research was awarded through the Innovation Partnerships Programme of Enterprise Ireland, which “encourages Irish based companies to work with Irish colleges to access their expertise and resources to develop new and improved products, processes, services, and generate new knowledge and know-how”.<sup>1</sup> This work is supported jointly by the Symantec Corporation

---

<sup>1</sup> Enterprise Ireland:  
<http://www.enterprise-ireland.com/en/Funding-Supports/Company/Establish-SME-Funding/Innovation-Partnerships.html> [last visited 2010-10-05]

(Ireland)<sup>2</sup> and the School of Applied Language and Intercultural Studies (SALIS) of Dublin City University (DCU).<sup>3</sup>

Symantec is now one of the world's leading software companies, one well-known product of which is the Norton Antivirus. Symantec produces products on security, storage and systems management and provides services for individual consumers, small businesses and large enterprises. Headquartered in the U.S.A, Symantec localises its products globally. One of the key aspects of localisation is the translation of documentation which is usually written in English. The localisation department covering the EMEA (Europe, Middle East and Africa) area is based in Dublin. This group is in charge of translating the original English documents into roughly 30 target languages including French, Spanish, etc. With the emerging economic development, China has become a country with a strong purchasing ability and also has one of the biggest communities of online users (Spethman et al. 2009). It is one of the main markets where big software and other companies, such as Symantec, Microsoft and IBM, localise their products. Given the size of the Asian market and due to the fact that, compared to most European countries, English is less understood in China, translating documents into Chinese accounts for a significant volume of translation every year.

This background motivates the focus of this study, i.e. translation of documents in the IT domain from English into Chinese (simplified Chinese characters). Being a native speaker of Chinese allows the author to analyse the translations competently and thoroughly.

---

<sup>2</sup> Symantec: [www.symantec.com](http://www.symantec.com) [last visited 2010-10-05]

<sup>3</sup> SALIS, DCU: <http://www.dcu.ie/salis/index.shtml> [last visited 2010-10-05]



In the MT research community, it is more common to study translation into English than from English to other languages. However, the situation is different in a localisation environment where most documents are written in English and need to be translated into other languages. Hence, the findings of the study can be generalised into other localisation contexts.

Translation technologies such as MT and Translation Memory (TM) technology have been employed by Symantec's localisation team to translate product documentation for several years. Hutchins pointed out that "Machine translation is demonstrably cost-effective for large scale and/or rapid translation of technical documents and software localisation materials" (2002: 159). Likewise, Roturier mentioned that the deployment of translation technologies has "increased the speed and consistency of translation, which in turn also enabled us [Symantec] to process larger document sets within shorter turnaround times" (2009: 1). The basic translation process in Symantec is first to translate a document by the in-house MT system (a commercial customised rule-based system Systran) and then to post-edit the raw MT translation using external vendors (Roturier 2009).

Systran is a widely used commercial rule-based MT system which is utilised also by other companies and institutions such as CISCO,<sup>4</sup> EADS<sup>5</sup> and the European Union. In addition, as technical documents are commonly translated by localisation companies we argue that the general approaches we propose in this study can be beneficial to other organisations as well.

We are privileged in many ways through this industry-academia collaboration. As well illustrated by Aranberri (2009), the advantages of this type of

---

<sup>4</sup> CISCO: <http://www.cisco.com/> [last visited 2010-10-05]

<sup>5</sup> EADS (European Aeronautic Defence and Space Company): <http://www.eads.com/eads/int/en.html> [last visited 2010-10-05]

collaboration include access to funding, cutting-edge technologies, rich data sets and human resources. In addition, we also have access to reports from real users of MT technology about the core problems affecting translation quality. Our research question arose through examining the errors in the Chinese translation of IT documentation generated by Systran and recorded by Symantec. On the other hand, this collaboration is not without tensions due to the different objectives and cultures of the two parties, i.e. research for public science versus for commercial benefits (Carpenter et al. 2004). Moreover, the funding or resources provided may be limited to a certain extent. However, once agreement is reached by the two parties to respect mutual interest, this collaboration is deemed to be of great benefit to both sides (Carpenter et al. 2004; Aranberri 2009).

## **1.2 Research Questions**

Machine translation of prepositions has been recognised as a problem by many researchers, such as Wu et al. (2006), Li et al. (2005) and Flanagan (1994). However, as mentioned earlier, most research focuses on the translation of Chinese into English instead of the reverse direction which is more relevant to localisation scenarios. This study contributes to the knowledge about translation of prepositions by MT systems by answering the following questions. Firstly, which prepositions are translated unsatisfactorily (i.e. human post-editing is required)? Secondly, which errors occur most frequently in our corpus? Thirdly, what are the most salient errors associated with each preposition? Fourthly, what existing solutions are suitable for tackling the most common errors? Finally, what solutions are there that have not yet been tested and, of these which are the most effective? The detailed contextualisation of the research questions will be presented in Chapter 3.

### **1.3 Structure of the Thesis**

The chapters which follow are dedicated to addressing the above research questions. Chapter 2 begins by giving a brief overview of the state-of-the-art MT systems involved in this study, including rule-based and data-driven systems. One rule-based and one data-driven system were employed in this study, i.e. Systran (a rule-based system) and Moses (a data-driven system). We then introduce approaches to MT evaluation which play an indispensable role in the development of machine translation. The evaluation of the quality of a machine translation can be conducted manually by human evaluators or automatically by automatic metrics. Due to their respective shortcomings and benefits, both approaches to evaluation are often conducted at the same time in a study.

Chapter 3 contextualises the research questions and provides a concise discussion of the characteristics of Chinese and English prepositions. An error typology of prepositions is set up for further examination. In addition, existing methods that have been proposed to enhance the performance of the RBMT system are also reviewed.

Chapter 4 describes the methodology employed in this study. The corpora are prepared following the principles in the literature to address issues affecting validity and reliability in order to obtain valid results. The rationale of the selection of MT systems and evaluation approaches is presented. An exploratory pilot project testing several existing methods for improving translation of prepositions is conducted. The results set a path for our further investigation of the most promising methods.

Chapter 5 presents the first human evaluation which attempts to answer the first three research questions. By examining the errors in the translation of a test

sample, information such as how many prepositions are translated problematically and what errors are associated with each preposition can be revealed. Based on this information, several approaches are proposed to reduce the errors.

Chapter 6 and Chapter 7 introduce the main approaches we propose for improving the translation quality of prepositions. The procedure for conducting statistical post-editing, especially the procedure for preposition-targeted statistical post-editing is introduced in Chapter 6.<sup>6</sup> Statistical source re-writing and automatic dictionary extraction are discussed in Chapter 7.<sup>7</sup> In both chapters, both qualitative and quantitative results are reported. In addition, results of human evaluation are accompanied by the results of automatic evaluation metrics to ensure the reliability of the results. The correlation between human evaluation and automatic evaluation is scrutinised and a new evaluation methodology is proposed.<sup>8</sup>

Chapter 8 reports a comparison of the translation from four different systems with a particular focus on their translations of prepositions. The rationale and preparation of the comparison is introduced first, followed by a detailed linguistic analysis of the translations. Finally, the last chapter summarises the findings and lessons learnt from this research. Additionally, directions for future research are identified.

---

<sup>6</sup> Part of the work has been discussed in *A comparison of statistical post-editing on Chinese and Japanese* (Tatsumi and Sun 2008) published in the *International Journal of Localisation*.

<sup>7</sup> This work has been introduced in *A novel pre-processing method for a Rule-Based Machine Translation system* (Sun et al. 2010) presented at the 14th Annual Conference of the European Association for Machine Translation (EAMT).

<sup>8</sup> Part of the work has been presented in *Mining the correlation between human and automatic evaluation at sentence level* (Sun 2010) presented at the 7th International Conference on Language Resources and Evaluation (LREC).

## **Chapter 2: Machine Translation and MT Evaluation**

Machine Translation (MT) in general is “the now traditional and standard name for computerised systems responsible for the production of translations from one natural language into another, with or without human assistance” (Hutchins and Somers 1992: 3). MT has come a long way since the idea was first outlined in Warren Weaver's historical memorandum in 1949 (Hutchins 2000). Broadly speaking, there are two types of MT architectures at the moment: Rule-Based MT (RBMT) and Statistical MT (SMT). Meanwhile, a new trend - hybrid MT - has become increasingly popular recently (Carl and Way 2003). Rapid development of MT systems is predicated on rapid evaluation of MT outputs. In return, evaluation stimulates further improvement of MT. In this chapter, a brief review of MT and evaluation of MT is presented.

### **2.1 Machine Translation**

We begin by looking at RBMT in Section 2.1.1 as the major MT system used in this study pertains to this group. An RBMT system is based on manually crafted linguistic rules. The performance of the system relies heavily on the coverage and quality of these rules. Therefore, linguistic information plays a crucial role in the process of system development. However, with the increase of freely accessible data, highly automated data-driven approaches have become the dominant MT architecture. SMT, as one of the data-driven approaches in particular, is dominant in the research area at the moment. Such MT systems require less human labour than traditional Rule-Based approaches. SMT was first introduced by Brown et al. (1990) and its mathematical foundation was detailed by Brown et al. (1993). It has evolved from basic word-based models into complex phrase-based models and

now into more sophisticated linguistics-rich hierarchical models. Detailed information on building and deploying an SMT system can be found in Section 2.1.2. With the steady development of both RBMT and SMT, recent research has included a hybrid MT approach (González et al. 2006; Habash 2002), where the advantages of different MT architectures can be combined into a more efficient framework. A description of these approaches can be found in Section 2.1.3.

### **2.1.1 Rule-based Machine Translation**

A classical RBMT architecture was first depicted by the well-known Vauquois Triangle (Vauquois 1968). Three approaches to RBMT evolved. The first (and the earliest) approach is generally referred to as the Direct approach (Hutchins and Somers 1992). An MT system in this approach is designed specifically for one particular pair of languages in one direction. A bilingual dictionary is needed to substitute the words in the source language (SL) with the corresponding equivalents in the target language (TL) (Arnold et al. 1994). This approach over-simplified the translation process in that it focuses on word-for-word substitution. The second approach is the Interlingua approach, which assumes the possibility of building pivot “meaning” representations (or Interlingua) (Hutchins and Somers 1992). Such representations are common to more than one language. Languages can translate to or from these representations. Translation is thus conducted in two steps: from SL into the Interlingua, and from the Interlingua into the TL. However, obtaining such an abstract and language-independent Interlingua can be quite challenging (Hutchins and Somers 1992). Therefore, a less ambitious approach – the Transfer approach – was developed which translates in three steps, namely, Analysis, Transfer and Generation. The transfer-based approach turns out to be the most practical approach and has become the most

widely used RBMT approach up to now (Hutchins and Somers 1992). As the RBMT system employed in this study belongs to this approach, Section 2.1.1.1 below reviews the three steps in more detail and Section 2.1.1.2 discusses today's Systran as an example of a Transfer-Based RBMT system.

#### **2.1.1.1 Transfer-based Approach**

A standard Transfer system usually consists of three steps as mentioned above: Analysis, Transfer and Generation. Analysis is the process of examining a SL sentence, changing it into a SL representation with most source ambiguities (such as words with more than one part of speech and ambiguous attachment structures) resolved using the linguistic rules of the SL. Analysis needs various aspects of monolingual linguistic knowledge including morphology, syntax and semantics (Hutchins and Somers 1992). Morphological analysis is generally considered as the first step of analysis. It can help to simplify the more difficult problems of syntactic and semantic analysis. Morphological analysis is reasonably straightforward for languages such as English and French; while for languages such as Chinese and Japanese, where the word boundaries are not orthographically marked, the morphological analysis can only be performed after the process of identifying the words, i.e. word segmentation (dividing sentences into meaningful units) (Sproat et al. 1996). We will come back to this in more detail when introducing SMT in Section 2.1.2.

After Analysis, the second step is Transfer. The necessity of this component is due to the lexical and structural differences between the SL and TL. Therefore, there are two major levels of transfer: lexical transfer and structural transfer (Hutchins and Somers 1992). Both processes have to face some challenges because of the diversity of languages. Lexical transfer ambiguities arise when a

single source language word can potentially be translated by a number of different target language words or expressions. Transfer ambiguities at structural level are even harder to capture and resolve compared to those at lexical level (Hutchins and Somers 1992).

Generation (also known as Synthesis) refers to the final stage of a transfer system, that is, the production of the final translation. In a transfer-based system, the generation phase is generally split into two modules: syntactic generation and morphological generation (Hutchins and Somers 1992). After the analysis and transfer stages, intermediate representations of both SL and TL are produced. These representations are now reassembled and converted into an ordered surface-structure tree with appropriate target language grammars and features. This tree then goes through morphological generation, generating lexical items in the target language.

As mentioned above, the transfer-based RBMT system is currently the most well-established system. Hence, it has become a common practice to simply use RBMT instead of Transfer-Based RBMT to refer to this type of system. Figure 2.1 shows the flowchart of an RBMT system and concludes this section.

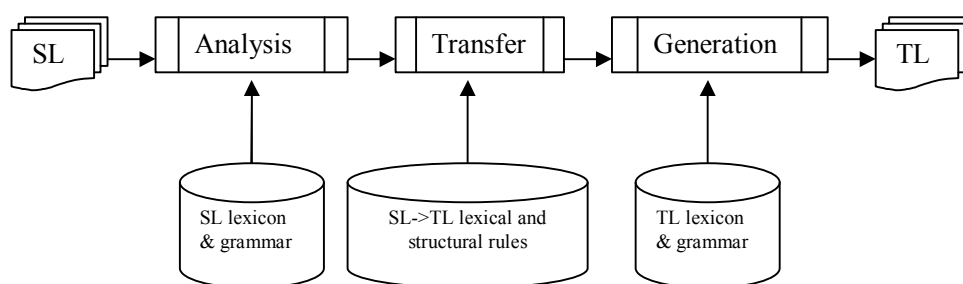


Figure 2.1: Flowchart of an RBMT system



### **2.1.1.2 Systran**

A number of RBMT systems were reviewed by Hutchins and Somers (1992) including the main MT system employed in this study – Systran.<sup>9</sup> Systran started off based on the Direct approach but later evolved into a Transfer-based system. It is currently widely used by industry as well as the research community. Being a Transfer-Based RBMT system today, the three basic translation steps discussed above also apply to Systran, namely, Analysis, Transfer and Generation (Senellart 2007; Attnäs et al. 2005).

The Analysis step is the first and the most important step of Systran (Surcin et al. 2007). Two types of analysis are involved: a Global Document Analysis (which identifies the subject domain) and a Grammatical Analysis (which performs a grammatical analysis and provides the system with the data required to represent the source language) (Senellart 2007). Information on part-of-speech, clause dependencies and relationships between entities of the sentence, as well as their functions, is extracted through the following modules: morphological analysis, grammatical disambiguation, clause identification and basic local relationships (Senellart 2007).

Based on the above obtained information, at transfer stage, the system attempts to transfer the source language sentence into the target language sentence in terms of structures and lexicon. It is the only stage where both the source and target languages are involved and described. In the last step, the generation model finishes the description of the target translation, removing unnecessary information and generating a translation that is as grammatical as possible.

---

<sup>9</sup> Systran's official website: <http://www.systran.co.uk> [last visited 2010-06-02]

All three steps depend heavily on linguistic resources created by experts: rules and dictionaries, both monolingual and bilingual. In an early description of Systran, Hutchins and Somers (1992) pointed out that the main components of Systran are the large bilingual dictionaries which not only provide “lexical equivalences but also grammatical and semantic information used during analysis and generation” (1992: 177). Arnold et al. (1994) also mentioned that, in terms of the amount of information provided, dictionaries are the largest components of an RBMT system. The size and the quality of the dictionary determine the coverage of an RBMT system and hence determine the quality of its translation (Gerber and Yang 1997). To date, dictionaries are still one of the most important parts of Systran with around 20 main dictionaries and other integrated dictionaries (Senellart 2007).

Apart from the built-in dictionaries, most commercial RBMT systems provide their users with a function for building their own domain-specific dictionaries (known as a User Dictionary (UD)). Dictionary entries have long been the main area for customisation of Systran by users (Dugast et al. 2009). The advantage of this function is that “the end users can expect to be able to contribute most to a system...to make some additions to the system ...[in order] to make the system really useful” (Arnold et al. 1994: 87).

With the development of computer technologies and new research in linguistics, various approaches have been proposed in order to improve the performance of Systran (and other RBMT systems). Antonopoulou (1998) discussed ways to resolve multiple meaning ambiguities in order to improve the performance of Systran. Building an RBMT system for a new language pair requires a large amount of effort; however, Surcin et al. (2007) explained how

Systran achieved its rapid development of new language pairs by reusing language rules. In terms of building up dictionaries, Dugast et al. (2009) proposed to craft lexical dictionaries semi-automatically with the use of parallel corpora. In addition, Dugast et al. (2007) and Simard et al. (2007a; 2007b) described how hybrid methods could greatly improve the performance of Systran.

An RBMT system, as just discussed, uses manually coded rules from experts to control the translation from one language to another. Therefore, the translations are syntactically correct, in general. However, the handling of complex sentences often fails and the engine generates incomprehensible translations if no rules for these complex sentences are encoded in the engine. Moreover, not having access to the grammatical rules of the system, the users cannot address this problem directly by modifying the system's rules. Lexical translation can be tuned to a certain domain. But accurate translation depends heavily on the dictionary coverage. The main downside of RBMT systems is that a large amount of time and human effort are required to develop and maintain the systems. Rules and dictionaries have to be manually crafted and manually validated and updated. In contrast, data-driven approaches acquire such knowledge automatically from large bilingual corpora. The next section describes two basic paradigms of the data-driven approach, with the dominant system (SMT) described in more detail, given its relevance to this study.

## **2.1.2 Data-driven Machine Translation**

### **2.1.2.1 Example-based Machine Translation**

The first type of data-driven approach is the so called Example-based Machine Translation (EBMT) (Arnold et al. 1994; Nagao 1984). To put it simply, an EBMT system performs three distinct processes in order to transform an SL sentence into a TL translation (Groves and Way 2005: 306):

- (1) Searching for similar matches (sentences or segments) at the source side of the bilingual corpora and obtaining their translations;
- (2) Determining the relationship between the retrieved segments;
- (3) Recombining the segments of the target translation to produce the final translation.

Many EBMT methods have been put forward. The difference between them lies in their matching criteria for “closest match” or “best match” (Koehn 2003). Overall, the quality of translations from an EBMT system increases with more stored translation examples (Carl and Way 2003). A more detailed review on EBMT systems can be found in Carl and Way (2003) and Somers (1999).

Translation Memory (TM), which is closely related to EBMT, has been widely used in the localisation area. A common feature shared by EBMT and TM is their use of a database of existing translations (Somers and Fernández Díaz 2004). A TM is a database of already-translated examples with both SL sentences and their corresponding TL translations either from human translation directly or from human edited MT output. A TM tool automatically compares a given SL sentence against the ones already stored in a TM and presents matches as translation suggestions based on the level of matches for human translators to work on.

However, unlike an EBMT system, a TM tool does not generate translations for sentences not matched in the database. One of the well-known TM tools is SDL Trados Translator's Workbench.<sup>10</sup> The corpora in our study are from an in-house translation memory of Symantec stored using this tool. We refer the readers to Somers and Fernández Díaz (2004) for a detailed comparison between EBMT and TM. A more practical introduction to various TM tools can be found in a user report by Lagoudaki (2006).

### **2.1.2.2 Statistical Machine Translation**

The second architecture within the data-driven approach is Statistical Machine Translation (SMT). Recently, SMT has become the predominant paradigm in the research area (Koehn 2010; Way 2010). In this section, the architecture of an SMT system is described in more detail followed by an example of the SMT system used in this study.

SMT is a “purely statistical and language-independent approach” developed from a “mathematical theory of probability distribution and probability estimation” (Carl and Way 2003: xix). SMT has developed from word-based models into more complicated phrase-based models and then recently into even more complicated syntactically-rich hierarchical models. A word-based SMT model mainly focuses on lexical translation, i.e. the translation of words in isolation by implementing a lexical translation probability distribution (Koehn 2010). The word-based SMT model is no longer the state-of-the-art model as it has been replaced by the Phrase-Based SMT (PB-SMT) model which is currently the best performing or leading paradigm, especially in the research field (Koehn 2010; Way 2010; Koehn 2003). PB-SMT models are based on the translation of

---

<sup>10</sup> SDL's website: <http://www.trados.com/en/> [last visited 2010-06-03]

word sequences instead of one word at a time (Koehn 2010). The two key notions involved are the translation model and the language model (or reordering model) (Koehn 2010; Arnold et al. 1994). The translation model consists of a bilingual phrase table with frequencies of phrases, knowledge of sentence lengths and relative positions of source and target words. The language model provides the SMT system with knowledge of the target language so that translations of a new text will be as grammatical as possible. We will explain how to obtain the translation and language models in more detail taking Moses (which is the SMT system used in this study) as an example.<sup>11</sup>

### **2.1.2.3 Moses**

Moses is an open-source MT toolkit as well as a stand-alone SMT system. It is by far the most popularly downloaded and accessed MT software according to TAUS (2010).<sup>12</sup> The official website of Moses lists a detailed step-by-step guide for installing, training, tuning and decoding (or translating). Building and using a Moses MT system consists of four basic steps: pre-processing, system training, system tuning (or evaluation) and decoding (in other words, translating).

As a data-driven MT system, the prerequisite for the training of any SMT system is a large amount of parallel bilingual corpora aligned at sentence level (Way 2010). The SL corpus contains sentences in one language (say, English) and the TL corpus contains translations in another language (Chinese, for example). The TL translations can either be direct human translation of the SL sentences, or MT output which is post-edited by humans. This human translation or human post-edited translation is usually called the *reference* translation by researchers.

---

<sup>11</sup> Moses' main page: <http://www.statmt.org/moses/> [last visited 2010-06-04]

<sup>12</sup> Translation Automation User Society (TAUS): <http://www.translationautomation.com/> [last visited 2010-04-28]

While the output quality from an RBMT system is strongly influenced by the coverage of language rules and dictionary entries, it is the size and the quality of the training corpora that influence the quality of an SMT system. According to Way (2010), systems are usually trained on several million words of data in order to achieve good translation quality. Large parallel corpora are made available by either language resource centres such as the Linguistic Data Consortium (LDC)<sup>13</sup> or large-scale academic projects such as EuroMatrix<sup>14</sup> and the Workshop on Machine Translation (WMT)<sup>15</sup> (see Section 2.2.4) or international institutions or governments as is the case for the EuroParl corpus (Way 2010).<sup>16</sup> Generally speaking, the bigger the corpora the better the translation. However, the quality of an SMT system is also decided by other factors such as the types of corpora used. Recently, more attention has been given to using the right training data or exploiting the full potential of existing data (Lü et al. 2007). Schwenk et al. (2009) found that adding extra out-of-domain corpora (one of which contained an additional 575 million English words) failed to achieve any improvement in their SMT system. The experiment conducted by Ozdowska and Way (2009) also clearly showed that quantity of training data is not always the only factor influencing the performance of an SMT system. However, to date, there is no standard agreement on the most suitable training data for a system.

Corpora are not created with MT in mind, and they have to be pre-processed before being used to train an SMT system. Parallel corpora have to be encoded conforming to the requirement of the tools used. Only data in plain text format can be used for training a Moses system at the moment. Therefore, formatted data

---

<sup>13</sup> Linguistic Data Consortium: <http://www ldc.upenn.edu/> [last visited 2010-06-07]

<sup>14</sup> EuroMatrix: <http://www.euromatrix.net/> [last visited 2010-06-07]

<sup>15</sup> WMT 2009: <http://www.statmt.org/wmt09/> [last visited 2010-06-07]

<sup>16</sup> EuroParl: <http://www.iccs.inf.ed.ac.uk/~pkoe hn/publications/europarl/> [last visited 2010-06-07]

have to be transformed into plain text format. In addition, the SL corpus and the TL corpus have to be aligned at sentence level.

Among all steps, tokenisation is one of the most important. Tokenisation is a process of dividing the sentences into white space-separated tokens (words and punctuation marks). As stated in Section 2.1.1.1, the division can be done comparatively easily for some languages such as English but is more difficult for other languages such as Chinese or Japanese whose word boundaries are not orthographically marked. Three of the main standards in Chinese word segmentation are the PRC (People’s Republic of China) national standard, the PKU standard (put forward by Peking University) and the Penn Chinese Treebank standard (Xia 2000). Many segmenters based on the standards have been created, some of the open-source segmenters are the LDC Chinese segmenter,<sup>17</sup> the Stanford Chinese segmenter,<sup>18</sup> the ICTCLAS Chinese segmenter.<sup>19</sup> In some cases, one can also modify a standard to meet his/her own purpose. For example, although mainly based on the PRC standard, Systran modified this standard for their own purposes (Yang et al. 2003). Table 2.1 shows an example of divergence in Chinese word segmentation (Yang et al. 2003: 180).

	中华人民共和国 (People’s Republic of China)	第一 (the first)	李白 (Li, Bai) <sup>20</sup>
PRC standard	中华人民共和国	第一	李 白
Systran’s segmenter	中华 人 民 共 和 国	第 一	李白
Symantec’s segmenter	中华人民共和国	第一	李白

Table 2.1: Divergences of Chinese segmentation

<sup>17</sup> LDC Chinese segmenter: [http://projects.ldc.upenn.edu/Chinese/LDC\\_ch.htm#cseg](http://projects.ldc.upenn.edu/Chinese/LDC_ch.htm#cseg) [last visited 2010-06-09]

<sup>18</sup> Stanford Chinese segmenter: <http://nlp.stanford.edu/software/segmenter.shtml> [last visited 2010-06-09]

<sup>19</sup> ICTCLAS Chinese segmenter: [http://www.nlp.org.cn/categories/default.php?cat\\_id=12](http://www.nlp.org.cn/categories/default.php?cat_id=12) [last visited 2010-06-09]

<sup>20</sup> A person’s name.



The examples show that “People’s Republic of China” is regarded as one Chinese word according to the PRC standard. The segmenter of Systran treated it as three Chinese words while the segmenter of Symantec treated it as one word. English does not have this problem as the words are already separated by white space.<sup>21</sup>

This study does not intend to explore the best segmentation standard for Chinese. What is important for a study is to keep the segmentation consistent in order to ensure valid comparisons. Therefore, the same segmenter should be used for all the corpora of the same language. Throughout this study, the segmenter used in Symantec is employed.<sup>22</sup>

Once the data sets are pre-processed, the monolingual target corpus is employed to build the language model through the language model toolkit which is used in most state-of-the-art SMT systems. Obtaining the translation model using the bilingual parallel corpora is also straightforward. The Moses toolkits have simplified the process of generating and deploying a statistical engine. There are two fundamental elements in system training: word alignment (word-to-word mapping between source language and target language) and phrase table extraction. High quality word alignment is essential for high quality output of SMT as phrases are extracted from word alignment tables which will be used at the decoding stage to generate translation options for a language pair (Ma 2009). The phrase table generated is like a bilingual dictionary. However, the term “phrase” is not used in the traditional grammatical sense. Rather, they are bilingual sequences with various numbers of words.

---

<sup>21</sup> Note that *People’s* also has to be separated into two tokens, i.e. *people* + *’s*.

<sup>22</sup> The segmenter is based on the PRC standard. Information about this segmenter can be found: <http://www.mandarin-tools.com/> [last visited 2010-06-09]

We will now give a detailed explanation of the content of the phrase table because it is of great importance to us at a later stage (see Chapter 6). Figure 2.2 is a snippet of an entry (the SL part is English and TL part is Chinese) from one Moses phrase table.

---

Lend me some money ? ||| 借 我 些 钱 吗 ? ||| (0) (1) (2) (3) (4,5) ||| (0) (1)  
 (2) (3) (4) (4) ||| 1 - 0.00163507 - 1 - 0.0189709 - 2.718

---

Figure 2.2: An entry from a Moses phrase table

The left most column is the source phrase, followed by the target phrase segmented into words. The numerals in the parenthesis indicate word alignment information (or the position of words) in both directions. The word alignment from English (SL) to Chinese (TL) is presented first then followed by the alignment from Chinese (TL) to English (SL). 0 represents the first word in a phrase and so on. Therefore, the first group of numbers (0) (1) (2) (3) (4, 5) indicate the position of the corresponding TL words. To put it in plain words: the first SL word aligns with the first TL word (0) and so on. Note that the last SL token (the question mark) aligns with the last two (the 4<sup>th</sup> and 5<sup>th</sup>) TL words, hence, both TL tokens are put in the same pair of parenthesis. Figure 2.3 shows this corresponding relationship from SL to TL in detail.

English	→	Chinese
lend	→	借 (0)
me	→	我 (1)
some	→	些 (2)
money	→	钱 (3)
?	→	吗 & ? (4, 5)

Figure 2.3: Word alignment information from SL to TL (0 represents the first word)

The second group of numbers: (0) (1) (2) (3) (4) (4) indicates the position of the corresponding SL words with regard to the TL words. This means that the first

TL word aligns with the first SL word (0) and so on. Again, since both the fourth and the fifth words at the TL side align with the fourth token (the question mark) at the SL side, (4) is repeated. Figure 2.4 shows the break-down of this alignment information.

<b>Chinese</b>	<b>→</b>	<b>English</b>
借	→	lend (0)
我	→	me (1)
些	→	some (2)
钱	→	money (3)
吗	→	? (4)
?	→	? (4)

Figure 2.4: Word alignment information from TL to SL (0 represents the first word)

Following the word alignment information in Figure 2.2 there are five numbers (1, 0.00163507, etc.) which represent the translation probabilities of the two phrases. The first number (“1”) is the phrase level probability that the whole TL phrase is the corresponding translation of the whole SL phrase; the third number (“1”) represents the reverse order. The second (“0.00163507”) and fourth (“0.0189709”) number are the average lexical level probabilities. The final number (“2.718”) is a default fixed phrase penalty value in Moses which prevents the translation from getting too long or too short.

After obtaining the translation and language models, the SMT system is ready to be used. One remaining option before putting the SMT system into use is to tune or optimise the system using new bilingual SL and TL corpora (tuning data) normally with a few hundred sentences. The new corpus should not contain sentences already existing in the training data. Moses uses Minimum Error Rate (MERT) (Och 2003) to tune the system by optimising BLEU (Bilingual Evaluation Understudy) scores (see Section 2.2.2.1). MERT estimates the best

parameters of the system with the new data through a series of iterations, trying to minimise the errors of the system and attempting to obtain the best BLEU scores. The validity of using BLEU for the fine-tuning of the system has been challenged (Zaidan and Callison-Burch 2009). However, Och and Ney (2003) supported this approach saying that tuning is necessary if BLEU is also used to evaluate the system's output after it has been built. Since BLEU is used in this study, MERT will be performed. Further information on Moses can be found on the official website of Moses and in Koehn et al. (2007).

SMT is a very different translation approach from RBMT as it does not require extensive linguistic knowledge. On the other hand, the approach is only suitable for languages with access to a large amount of bilingual parallel data. The fact that it can only handle plain text makes its use in many real life scenarios problematic, especially in production of rich formatted texts which are common for industry. In addition, the statistical nature could lead to unpredictable errors (Way 2010). Although the SMT approach has become the leading paradigm in the research field, most available commercial systems are still in the RBMT category.

MT evaluations show that different MT architectures have their unique pros and cons. SMT systems are robust and perform better in lexical selection compared to RBMT systems but RBMT systems perform better in word order than SMT systems (Thurmaier 2005). Chen and Chen (1996) summarised the advantages and disadvantages of the existing systems at that time. Based on this study, Eisele (2007) relisted the pros (marked by +) and cons (marked by -) of the three system types discussed above in Table 2.2.

	Syntax	Structural Semantics	Lexical Semantics	Lexical Adaptivity
Rule-Based MT	++	+	-	--
Statistical MT	--	--	+	++
Example-Based MT	-	--	-	++

Table 2.2: Pros and cons of SMT, RBMT and EBMT

The complementary individual strengths of the SMT and RBMT approaches suggest that a hybridisation approach might be beneficial. A brief introduction to related works on system hybridisation is presented in the next section.

### 2.1.3 Hybrid Machine Translation

Much current research in MT is neither based purely on linguistic knowledge nor on statistics, but includes some degree of hybridisation (Cavalli-Sforza and Lavie 2006, from Way 2010: 556).

Eisele (2007) defined two types of hybridisation of different types of MT systems: shallow integration or deep integration. Shallow integration simply integrates two or more systems into a larger system. Deep integration is a new paradigm that integrates the advantages of the two approaches together, either adding a statistical module for an RBMT system or adding syntactic constraints/rules to an SMT system (Eisele 2007). Way (2010) defined two groups: the multi-engine approach and integrated-system approach. A more fine-grained and illuminating categorisation of system hybridisation is found in Thurmair (2009) in which three basic categories are listed, i.e. system coupling, architecture extension and genuine hybrid architecture, which are further broken into sub-categories. In this study, the shallow integration and deep integration distinction are used since they represent a more general categorisation.

Shallow integration can be achieved in “a serial way” (Thurmair 2009: 2). An example is Statistical Post-editing (SPE) which has been studied by several researchers (Dugast et al. 2007; Simard et al. 2007a; 2007b). An SPE system for an RBMT system is built following the steps of training an SMT system. However, instead of feeding the SMT system bilingual corpora, monolingual aligned parallel corpora are needed: raw RBMT system outputs and corresponding reference translations both of which are in the target language. The engine learns the differences between a raw RBMT output and a reference translation, calculates the probabilities of the changes, and edits a new RBMT output based on the knowledge gained. Such combinations of RBMT and SPE systems are highly competitive when it comes to the final translation quality (Schwenk et al. 2009) with more grammatical output and increased lexical selection quality (Dugast et al. 2007), one of the weak points of pure RBMT systems.

Another common practice of shallow integration is done in “a parallel way whereby the best output is produced” from a number of MT systems (Thurmair 2009: 2). For example, Alegria et al. (2008) reported their approach to selecting the best output from three MT engines: an EBMT system, an SMT and an RBMT system. Mellebeek et al. (2006) reported a technique in which the input sentence was decomposed into smaller chunks and a translation was produced by combining the best chunks of translations from several MT systems, selected through a confidence score assigned to each MT system.

The second type of system hybridisation is deep integration. One way of achieving deep integration is through system extension:

System extension means that the system architecture basically follows the R[B]MT or SMT paradigm but is modified by including resources of the respective other approach. Modifications can occur as pre-editing

(i.e. the system data are pre-processed), or core modification (e.g. phrase tables are extended, dictionaries are enlarged etc. by the respective other approach) (Thurmaier 2009: 3).

For example, Dugast et al. (2009) reported their method of quickly obtaining an extra phrasal dictionary for their RBMT system through making use of the bilingual phrase table of an SMT system. The test results showed improvements in translation in terms of BLEU scores. In a reverse direction, Chen et al. (2007) incorporated phrases extracted from RBMT output into the phrase table of an SMT system. Eisele et al. (2008) also reported their method of combining systems through lexical resources. Recently, more challenging deep integration proposals have been put forward. Such proposals usually require programming and linguistic knowledge such as the work of Vandeghinste et al. (2008).

In the current study, we propose three new methods of combining an RBMT system and an SMT system. One of our methods is an attempt at deep integration similar to the work of Dugast et al. (2009) while the other two methods belong to the shallow integration group. More detail will be reported in Chapter 6 and Chapter 7 respectively.

## **2.2 Evaluation of Machine Translation**

It is widely recognised that evaluation plays an important role in the development of MT and language technologies in general. However, evaluation is a complex issue. In the area of MT, there are two types of commonly used evaluation methods: human evaluation and automatic evaluation. As the ultimate users of machine translation outputs are humans, human evaluation is regarded as the “gold standard” for machine translation. However, the labour-intensive (thus cost

implications) and highly subjective characteristics of human evaluation have led to the popularity of automatic evaluation metrics, such as BLEU (Papineni et al. 2002), Precision and Recall (Turian et al. 2003) and TER (Translation Edit Rate) (Snover et al. 2006) among others.

Evaluation can serve the following three general purposes: error analysis of systems; comparison of systems and optimisation of systems (Giménez 2009: 16). Since both human evaluation and automatic evaluation are conducted intensively in this study, this section reviews the benefits and drawbacks of both modes of evaluation.

### **2.2.1 Human Evaluation**

Several types of human evaluation have been defined. Hutchins and Somers (1992) pointed out that at different stages of the development of an MT system, there are different types of evaluation, including prototype evaluation, development evaluation, operational evaluation, translator evaluation and recipient evaluation. White (2003) summarised six types of evaluation: declarative evaluation, operational evaluation, feasibility evaluation, internal evaluation, usability evaluation and comparison evaluation. Each type of evaluation focuses on different issues and is normally conducted by different evaluators, e.g. researcher, developer and potential users.

The most famous (or infamous) and probably the first large-scale human evaluation on the quality of MT is the ALPAC (Automatic Language Processing Advisory Committee) report. Using humans as judges, the report described a study comparing the output of MT systems with outputs of human translators. The criteria employed included intelligibility and fidelity (Pierce et al. 1966). Intelligibility was measured on a 9-point (1 to 9) scale, from 1 being “hopelessly



unintelligible” to 9 “perfectly clear and intelligible” (Pierce et al. 1966: 69). Fidelity was measured on a 10-point (0 to 9) scale depending on how much information the translated output retained compared to the source sentence. Although the report resulted in severely reduced funding into MT research, the standards used in the ALPAC report had a great influence on many of the MT evaluations in the following years.

Later, another influential evaluation report was the Van Slype report (Van Slype 1979) on the performance of Systran at the European Commission (EC).<sup>23</sup> The original purpose of the report was to provide a comprehensive review of the existing methods of MT evaluation and to advise appropriate evaluation methodology for the EC. Ever since it was made publicly accessible in 2003, the attributes of the quality of a translation, e.g. comprehensibility, fluency, accuracy, have been a prototype framework for MT evaluation (King et al. 2003). These evaluation attributes were also used by the Defence Advanced Research Projects Agency (DARPA) in their evaluation projects attempting to create a methodology to evaluate several machine translation systems. They assessed 16 systems in total following pre-defined attributes of translations, i.e. informativeness, adequacy and fluency (White et al. 1994). Such attributes, especially adequacy and fluency, have become the standard methodology for DARPA and other large scale evaluation campaigns.

A common practice in MT evaluation is that evaluators design their own evaluation approach from scratch based on their own evaluation purpose and the systems involved, resulting in a lot of repetition (King et al. 2003). Therefore, some studies have tried to standardise or unify the evaluation process such as the

---

<sup>23</sup> Van Slype report: <http://www.issco.unige.ch/en/research/projects/isle/van-slype.pdf> [last visited 2010-06-11]

Expert Advisory Group on Language Engineering Standards (EAGLES) set up by the EC.<sup>24</sup> The purpose of EAGLES is to advocate guidelines or general requirements before carrying out an evaluation. A seven-step recipe was proposed by the EAGLES evaluation working group in order to carry out a successful evaluation of language technologies.<sup>25</sup>

The Japan Electronic Industry Development Association (JEIDA) published their evaluation methodologies tailored to different users of MT (Nomura and Isahara 1992). A set of criteria were devised and could be followed if one of the following three types of evaluation was carried out: user evaluation of economic factors (to decide whether MT should be introduced and which type of system would be the most economical one), technical evaluation by users (to find out which system would best fit the needs of the environment) and technical evaluation by developers (to check if an MT system meets the original objectives).

Building upon previous work, the Evaluation Working Group of the International Standard in Language Engineering (ISLE) project (1999-2002) extended the principles proposed.<sup>26</sup> They organised several workshops and developed a Framework for Machine Translation Evaluation (FEMTI).<sup>27</sup> FEMTI aims at helping evaluators to choose the appropriate metrics based on the intended context of use. The project however, does not put forward any new metrics, but aims to “build a coherent picture of the various features and metrics that have

---

<sup>24</sup> EAGLES online: <http://www.ilc.cnr.it/EAGLES/home.html> [last visited 2010-06-11]

<sup>25</sup> EAGLES’ 7-step recipe: <http://www.issco.unige.ch/en/research/projects/eagles/ewg99/7steps.html> [last visited 2010-06-11]

<sup>26</sup> ISLE: <http://www.issco.unige.ch/projects/isle/> [last visited 2010-06-12]

<sup>27</sup> FEMTI: <http://www.issco.unige.ch/femti> [last visited 2010-06-12]

been used in the past, to offer a common descriptive framework and vocabulary, and to unify the process of evaluation design” (Hovy et al. 2002: 44).

Two quality aspects widely used in the evaluation projects are: fluency (or intelligibility) and adequacy (or fidelity) (Flanagan 2009; Hovy et al. 2002). The LDC further discussed and optimised these two concepts (fluency and adequacy) for the annual NIST Machine Translation Evaluation (LDC 2005).<sup>28</sup> According to their definition, adequacy indicates how much of the meaning expressed in the reference is also expressed in a translation and fluency refers to how fluent the translation is (Callison-Burch et al. 2007). A five point scale (1-5) was deployed for both aspects. A brief interpretation of adequacy and fluency scores can be found in Table 2.3 (Callison-Burch et al. 2007).

Rating	Adequacy	Fluency
5	All	Flawless
4	Most	Good
3	Much	Non-native
2	Little	Disfluent
1	None	Incomprehensible

Table 2.3: Interpretation of fluency and accuracy scores

The problem of scoring is that even with clear guidelines at hand, human evaluators still found it hard to assign appropriate scores to a translation. In recent evaluation campaigns, ranking has become the mainstream evaluation measurement (Callison-Burch et al. 2009; 2008). Humans are asked to compare outputs from several systems (or from the system at different development stages) and rank the outputs from best to worst relative to other outputs of the same source sentence. Ranking is found to be quite intuitive and reliable according to

---

<sup>28</sup> LDC: <http://www ldc.upenn.edu/> [last visited 2010-06-12]

Vilar et al. (2007). Moreover, compared to assigning scores, ranking can simplify the decision procedures for human evaluators (Duh 2008).

All the above-mentioned human evaluation metrics focus on judging the quality of whole sentences or documents. In some cases, evaluation is required for certain structures or constituents of a sentence. A constituent-based evaluation was reported in the work of Callison-Burch et al. (2007) in which syntactic constituents were randomly selected from the source sentence and the translations were ranked. In this study, since we are particularly interested in the translation of prepositions, constituent-based evaluation will be employed and the syntactic constituents will focus on prepositions and prepositional phrases.

In other cases, the purpose of evaluation is not to obtain a general score but to focus on the errors an MT system makes. An error analysis can be of great help to MT researchers or developers in the sense that it can pinpoint the key area where an MT system can be improved. Errors related to the RBMT system and the translation of prepositions in this corpus will be introduced in Chapter 3.

Two more issues pertaining to human evaluation are: who the evaluators are and how many there are. Human evaluators can be experts (translators) or non-experts depending on the context and the resources available (Aranberri 2009). There are both advantages and disadvantages to using professional translators vs. non-translators. Professional translators can deliver a more reliable assessment but there are cost implications. On the other hand, it is comparatively easy to find non-expert volunteers but using them carries risks such as inconsistent or random assessment (Aranberri 2009), thus affecting the validity of the results. As to the adequate number of evaluators to use, Carroll (1966, cited in Pierce et al. 1966) concluded that at least three or four evaluators should be used.

Arnold et al. (1994) also mentioned that a minimum of four evaluators should be used and the more the better. To solve the difficulty of finding a large group of human evaluators while having a restricted budget, Zaidan and Callison-Burch (2009) explored the possibility of using an online marketplace (Amazon's Mechanical Turk). This method has been used and studied more and more lately.<sup>29</sup>

Other human evaluation measures commonly used include: reading time, post-editing time and cloze test (Giménez 2009; Dabbadie et al. 2002).

Human evaluation is not without problems. Giménez (2009) listed the following drawbacks of human evaluation:

- (1) Human evaluation is both labour-intensive and time-consuming;
- (2) Human evaluation is static and not reusable;
- (3) Human evaluation can be easily affected by human factors which are not readily controllable, i.e. human emotions and tiredness. What's worse, there is no "right translation". One sentence can be translated differently by different people and all might be considered acceptable.

Although these shortcomings of human evaluation have been agreed by many researchers (such as Callison-Burch et al. 2006; Coughlin 2003), there is no alternative that can replace the role of human evaluation completely. Human evaluation still plays an indispensable role in providing valuable information on the quality of MT systems.

---

<sup>29</sup> The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NACCL) organised a workshop on the use of Amazon's Mechanical Turk.

## 2.2.2 Automatic Evaluation

Along with the development of MT technology from RBMT to SMT, automated evaluation also emerged as a quick, cheap, consistent and language-independent evaluation metric for MT (Papineni et al. 2002). Automatic evaluation metrics have become an important component in the development cycle of an MT system. In this section, we will mainly introduce some of the most commonly used automatic metrics, such as BLEU,<sup>30</sup> GTM (General Text Matcher) (Turian et al. 2003),<sup>31</sup> TER.<sup>32</sup> A short discussion is also presented about other metrics such as WER (Nießen et al. 2000).

### 2.2.2.1 BLEU

The central idea behind BLEU is that the closer an MT output is to a standard (human) translation, the better it is (Papineni et al. 2002). It is a precision-based (modified precision, to be more precise) metric that compares a system's output against one or several reference translations by summing over the n-gram matches found (starting from unigrams (1 word) to bigrams (2 words) to trigrams (3 words) and so on) and then dividing by the sum of words found in the reference translation set (Way 2010). Let us illustrate how BLEU works through an example mentioned by Papineni et al. (2002: 312):

---

Example 2.1

**MT output:** the the the the the the the.

**Reference 1:** The cat is on the mat.

**Reference 2:** There is a cat on the mat.

---

---

<sup>30</sup> BLEU: <http://www.nist.gov/speech/tests/mt/2008/scoring.html> [last visited 2010-06-13]

<sup>31</sup> GTM: <http://nlp.cs.nyu.edu/GTM/> [last visited 2010-06-13]

<sup>32</sup> TER: <http://www.cs.umd.edu/~snoover/tercom/> [last visited 2010-06-13]

Standard unigram matching goes like this: match the MT output word by word against all of the references and check if a word of the output is present in the references or not.

In example 2.1, all the occurrences of *the* in the MT output can be matched in both references. In other words, the output is of very high precision. Obviously, this is a false high precision. The problem with this type of unigram matching is that a reference word is not considered exhausted after being matched to an MT output word. To avoid such cases, Papineni et al. (2002) used a modified precision method for BLEU. To calculate this, one has to compare the number of times an n-gram appears in the MT output and the number of times this n-gram is in the reference. One has to “truncate each word’s count, if necessary, to not exceed the largest count observed in any single reference for that word” (ibid: 312). In the above example, the output contains seven occurrences of *the*, however, the maximum frequency of *the* in each of the reference sentences is two. The modified unigram precision for the output is  $2/7$ . Bigram, trigram or 4-gram precisions are zero as in the references there are no occurrences of bigrams (for example, *the the*), trigrams (*the the the*) or 4-grams (*the the the the*). A modified precision for a whole text is calculated based on the n-grams precision for each sentence. In addition to the precision scores, a brevity penalty (*BP*), which takes the length of the output (*c*) into consideration, is also calculated. It rewards an MT output which has similar sentence length (*r*), word selection and word order as the reference sentences. The final BLEU score is the geometric mean of the n-grams’ modified precisions multiplied by the exponential brevity penalty factor (readers are referred to Papineni et al. 2002 for a more detailed explanation of the formulae (1) and (2)).

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

$$(2) BLEU = BP * \exp \left\{ \sum_{n=1}^N W_n \log p_n \right\}$$

Although the process can be applied to any n-grams, experiments showed that 4-grams correlate best with their monolingual human evaluation (ibid). Therefore, BLEU with 4-grams has become the default. For the above example, the BLEU score with 4-grams is zero.

Since there is not only one correct translation for a sentence, and BLEU was developed to work with multiple references, Recall (which computes how many words in the references co-occur in an MT output) is not included in BLEU (ibid: 315). Another characteristic of BLEU is that it was designed with document or system-level evaluation in mind. Although it is currently the most commonly used metric, it has been criticised for its inefficiency at sentence level (Callison-Burch et al. 2006).

#### 2.2.2.2 GTM

In an attempt to get better evaluation at sentence level, GTM, based on precision, recall and the F-measure is proposed (Turian et al. 2003). The main concepts behind GTM are “maximum matching” and the “maximum matching size” (MMS). Let us explain this using an example (example 2.2). All example sentences shown in this study are from the technical documents of Symantec and all MT outputs are from Systran, unless otherwise specified.

---

Example 2.2

**MT output:** About the proactive threat scans the detect processes.

**Reference:** About the processes that the proactive threat scans detect.

---



Following the practice of Turian et al. (2003), we present the matching relationship between the MT output and the reference in a bitext grid in Figure 2.5. The dots (or hits according to Turian et al. 2003) in Figure 2.5 indicate that identical words are shared by the MT output and the reference.

MT Output ↑	processes			●					
	detect								●
	the		●						
	scans							●	
	threat						●		
	proactive					●			
	the		●						
	About	●							
		About	the	processes	that	proactive	threat	scans	detect
		Reference →							

Figure 2.5: (GTM) Bitext grid between an MT output and the reference

In this example all words in the MT output can be matched in the reference while in fact, *the* in the reference was double-counted. There is only one *the* in the reference but two in the MT output. To overcome this, the concept of “maximum matching” was employed. A matching only counts words in common between an output and a reference without allowing double-counting. In example 2.2, the output has eight words with only seven matching with the reference. A maximum matching refers to a block of maximum number of matched words between an output and the reference translation. There are four maximum matchings marked by the cells in grey. In order to take word order into consideration, rewards are assigned to longer matches through a special weighting. The size of maximum matching (MMS) is calculated using the following formula (3):

$$(3) \text{ MMS} = \sqrt[e]{\sum_{r \in M} \text{length}(r)^e}$$

$M$  in the formula refers to a set of maximum matchings.  $r$  is the size of any maximum matching.  $e$  refers to the weight assigned to reward longer matches.

One can tune this weight to assign rewards appropriate to one's study. The default weight of GTM is 1, in other words, no word order penalty is assigned. Other weights (such as  $e=2$  or  $e=3$ , etc.) were also employed (Giménez et al. 2005; Turian et al. 2003). In Figure 2.5, there are four maximum matchings with 2, 3, 1 and 1 word respectively. If we assign  $e=2$  to the matching and replace the formula with real numbers, the formula then looks like this:

$$MMS = \sqrt[3]{(1^2 + 1^2 + 2^2 + 3^2)}$$

Next, dividing MMS by the length of the output (C) or the length of the reference translation (R) will get precision (4) and recall (5) respectively.

$$(4) \text{ Precision } (C|R) = \frac{MMS(C.R)}{|C|}$$

$$(5) \text{ Recall } (C|R) = \frac{MMS(C.R)}{|R|}$$

(Turian et al. 2003: 2)

Precision measures how many words produced in the output match the translation in the reference. Recall tells how many words in the reference have been generated also in the output. Besides precision and recall, their harmonic mean “F-measure” (van Rijsbergen 1979) is also calculated to represent the percentage of matching between the output and the reference. This method can be extended to calculate scores at document level. More information can be found in Turian et al. (2003) and Melamed et al. (2003).

### 2.2.2.3 TER

Another approach to MT evaluation metrics tries to measure the post-editing effort of a human to change an MT output into a reference translation. One

example of such a metric is Translation Edit Rate (TER) (Snover et al. 2006). It was defined as:

The minimum number of edits needed to change a hypothesis [a candidate MT output] so that it exactly matches one of the references, normalised by the average length of the references. (6)

$$(6) \text{ TER} = \frac{\text{\# of edits}}{\text{average \# of reference words}} \quad (\text{ibid: 225})$$

*edits* in formula (6) include insertions, deletions, and substitutions of single words and also shifts of word sequences. The penalties are the same for all edits. Snover et al. (2006) pointed out the similarities between TER and GTM, in that a word is only allowed to be matched once and both allow reordering. However, TER does not particularly reward longer matches as GTM does.

TER calculates the number of insertions, deletions, substitutions and shifts required to change an output into a reference translation. If an output is compared to multiple references, the lowest number of edits will be used (example 2.3).

---

#### Example 2.3

**MT output:** Tony Blair Put on President Mubarak New Ideas to Advance the Peace Process.

**Reference 1:** Tony Blair Proposes New Ideas to President Mubarak to Drive Peace Process Forward.

**Reference 2:** Tony Blair Puts Forward New Ideas to President Mubarak to Drive Peace Process.

**Reference 3:** Tony Blair Presents Mubarak with New Ideas to Move Peace Process.

**Reference 4:** Tony Blair Presented President Mubarak with New Ideas to Push Peace Process.

---

If we measure the output against all four references using TER, we can get the following summarised report:

---

Best ref: reference 2

Average word of all references: 15.5

Number of edits: 5 (0 insertion and deletion; 4 substitutions; 1 shift)

TER score:  $5/15.5 = 0.3226$

---

From the report we can see that to change the MT output into the second reference requires the least number of edits. The specific number of edits is 5. The final TER score is calculated by putting the number of edits into formula (6).

Unlike BLEU or GTM score which ranges from 0 to 1, TER has no upper bound on its score. If an output is perfectly matched with a reference (i.e. no post-editing, insertion, deletion, shift is needed, the score will be 0). To sum up, for GTM and BLEU, the higher the score, the better the translation; for TER, the reverse is true.

#### **2.2.2.4 Other Metrics**

All three metrics discussed are string-based (or lexical-based) metrics. The closeness between an output and the corresponding reference translation at surface level is measured. The downside of string-based metrics is that the acceptability of an output to a human is not fully indicated by the scores. Besides, string-based metrics have been found to favour the output of SMT systems over that of RBMT engines while human evaluations show a reverse preference (Callison-Burch et al. 2007; Coughlin 2003).

Effort has been put into combining more information into string-based metrics, such as HTER (Snover et al. 2006), TERp (Snover et al. 2009) and METEOR (Metric for Evaluation of Translation with Explicit Ordering) (Banerjee and Lavie 2005). Additional knowledge or information is needed to get the best

use of these metrics. HTER requires human translators to first post-edit the MT output into acceptable translations using as few changes as possible which will then function as the reference to which the original MT output will be scored. METEOR functions better with a database of synonyms, such as WordNet for English;<sup>33</sup> TERp requires sets of paraphrases which also function as “synonyms” of phrases.

Recently, some researchers employed syntactic structures and dependency information extracted through parsing the MT output and the reference to build automatic metrics (Owczarzak et al. 2007a; 2007b; Liu and Gildea 2005). Some use machine learning techniques (Albrecht and Hwa 2007; Russo-Lassner et al. 2005) in their evaluation approaches.

The problem with these metrics is either that they are time and labour consuming (such as HTER) or need extra linguistic information (such as TERp and Meteor). Such knowledge bases are easy to obtain for English or other European languages but scarce for other languages such as Chinese (the language evaluated in the current study). Therefore, in large scale MT evaluation campaigns where various language pairs are involved, string-based metrics have been constantly updated and widely used.

It is worth pointing out that novel evaluation metrics are constantly being put forward. For example, Doherty and O’Brien (2009) explored the use of an eye-tracker (hardware that records the movement of one’s eyes while one is reading text on screen) as a means of evaluation for MT output.

Automatic evaluation metrics are generally consistent and stable no matter when and by whom they are used. Compared to human evaluation, automatic

---

<sup>33</sup> WordNet for English: <http://wordnet.princeton.edu/> [last visited 2010-06-14]

evaluations are fast, less costly and objective (Giménez 2009). Automatic evaluation meets the requirement of instant evaluation during the development of a system. However, the scores reported are arguably not indicative of the absolute quality of MT but are a superficial comparison of the closeness between an output and a reference sentence at lexical level (e.g. string-based metrics). In addition, a set of references – either human translated or human post-edited translation – is needed for all automatic MT evaluation metrics. For example, although proponents of BLEU claim that its advantage is that it can measure the MT output against more references to reflect the real quality of the MT output, to produce more references is both time consuming and costly. The last but not least problem of automatic evaluation is that the reliability of the scores has to be verified by their correlation with human evaluation.

### **2.2.3 Meta-evaluation of Human and Automatic Evaluation**

The success of automatic evaluation metrics has to be determined by their correlation with human evaluation, i.e. whether the judgement of automatic metrics equals the opinion of humans. Depending on the type of human evaluation used, the correlation between automatic and human evaluation is usually measured by Pearson's correlation coefficient or Spearman's ranking correlation coefficient or consistency level (Callison-Burch et al. 2009; 2008; 2007). The correlation value ranges from -1 to 1 representing negative correlation to perfect positive correlation.

In the MT research community, most effort has been put into finding out which automatic metric correlates better with human evaluation at the corpus level. Nevertheless, increasing attention is being paid to correlation at sentence level. According to Lin and Och (2004), high sentence level correlation of

automatic and human evaluation is crucial for machine translation researchers. Russo-Lassner et al. (2005) also pointed out that automatic metrics of high sentence level correlation could “provide a finer-grained assessment of translation quality” and could also “guide MT system development by offering feedback on sentences that are particularly challenging” (ibid: 3). The correlation scores reflect how similar automatic metrics and human evaluation are in judging the quality of an MT output. However, this correlation varies with the languages evaluated, the type of documents tested and the system involved. For example, BLEU correlates better with human evaluation at document level than at sentence level (Turian et al. 2003). From her study, Aranberri (2009) found that for French, Japanese and German GTM correlated better with human evaluation than BLEU and TER.

Therefore, no concrete conclusion has been made so far as to what is the best automatic metric. Turian et al. (2003) concluded that automatic MT evaluation measures are far from being able to replace human evaluation, especially at sentence-level. An area less studied is how to best make use of both automatic and human evaluation. Sun (2010) reported on a pilot project which uses an automatic metric to increase the reliability of human evaluation. The findings can be helpful in two ways: first, it opens new ways of using automatic metrics to distinguish translations with a real difference in quality; second, it reduces the effort in human evaluation by selecting specific translations instead of all translations to be evaluated. More details will be reported in Chapter 6.

The gold standard of evaluation – human evaluation – is not without problems, either (cf. Section 2.2.1). In some cases, human judgements are not consistent with each other or even with themselves. To ensure the reliability of the human evaluation results, the inter-evaluator and intra-evaluator correlation has to

be examined before drawing conclusions from the results. A common measurement of the reliability of human evaluation is to calculate the correlation coefficient through Kappa statistics (Carletta 1996; Fleiss 1971). A Kappa coefficient (K) is calculated based on two levels of agreement, i.e. the observed agreement ( $P_{obs}$  how much agreement is actually present) and the expected agreement ( $P_{exp}$  how much agreement would be expected by chance) following the formula (7) below (Viera and Garrett 2005):

$$(7) \quad K = \frac{P_{obs} - P_{exp}}{1 - P_{exp}}$$

According to the definition set by Landis and Koch (1977), a Kappa score between: 0.0 - 0.20 signifies slight agreement; 0.21 - 0.40 signifies fair agreement; 0.41 - 0.60 signifies moderate agreement; 0.61 - 0.80 signifies substantial agreement; 0.81 - 1.00 signifies almost perfect agreement. Kappa values reported in MT research vary across studies. For example, the inter-evaluator correlations of ranking several MT outputs reported by Callison-Burch et al. (2009; 2008; 2007) are all fair agreements (0.323, 0.367, and 0.373 respectively). In her study of judging whether a translation is correct or not, Aranberri (2009) reported the inter-evaluator correlation ranges from no correlation to substantial agreement.

In summary, there are both benefits and drawbacks to human and automatic evaluation and, as a consequence, in order to obtain more valid evaluation results, both human and automatic evaluations (usually several automatic metrics at the same time) are employed in major evaluation campaigns.



## 2.2.4 MT Evaluation Campaigns

The MT systems and the evaluation metrics we have discussed above have been compared and reported in many large-scale evaluation campaigns, such as NIST<sup>34</sup> evaluation (which is supported by the National Institute of Standards and Technology of the U.S.A) and the TC-STAR project (which is financed by the European Commission).<sup>35</sup> In addition, IWSLT (International Workshop on Speech Language Translation),<sup>36</sup> WMT<sup>37</sup> (Workshop on Statistical Machine Translation) and CWMT (China Workshop on Machine Translation) are also popular evaluation campaigns in the area.<sup>38</sup>

In general, the purpose of large-scale evaluations is to present and hopefully advance the state-of-the-art of MT technologies and the state-of-the-art MT evaluation technologies. MT evaluation campaigns can be characterised into several categories depending on the criteria used. Some campaigns focus on speech translation such as IWSLT and TC-STAR while some focus on text translation such as NIST and SMT workshop evaluations. MT evaluations can also be separated according to the language pairs being translated. The main language pairs in which NIST is interested are Arabic to English and Chinese to English (the 2009 NIST also included Urdu to English translation). While WMT studies translations between European language pairs, CWMT calls for participation on Chinese to and from English translation and Chinese to Mongolian translation. For these languages, as we explained above, string-based

---

<sup>34</sup> NIST open Machine Translation Evaluation: <http://www.nist.gov/speech/tests/mt/> [last visited 2010-04-28]

<sup>35</sup> TC-Star Machine Evaluation: <http://www.tc-star.org/> [last visited 2010-04-28]

<sup>36</sup> IWSLT Machine Evaluation in 2010: <http://iwslt2010.fbk.eu/node/15> [last visited 2010-04-28]

<sup>37</sup> WMT Workshops: <http://www.statmt.org/> [last visited 2010-04-28]

<sup>38</sup> CWMT Evaluation: <http://www.icip.org.cn/cwmt2009> [last visited 2010-04-28]

evaluation metrics are considered as the most suitable metrics and hence are commonly used.

Depending on campaigns and designs of MT systems, an MT system may only show up in certain types of comparisons or appear in many different types of evaluations. The best system in each campaign varies depending on language pair, method of evaluation and training and test corpus employed. Among all the participating systems, Google and Systran, which all have their online versions available for general users, are often among the list of the best systems. Callison-Burch et al. (2007) summarised the results of the 2007 WMT and found that Systran was greatly favoured by human evaluation by being ranked as the best system most often. The NIST 2008 official evaluation results showed that Google's SMT system (Google for short henceforth) was the best system for English to Chinese translation. The WMT-2009 results also showed that Google was always among the best systems for many language pairs (Callison-Burch et al. 2009).<sup>39</sup>

Similar smaller-scale evaluations are reported in the literature, too. For example, after comparing an SMT (Portage), an RBMT (Systran) and an SPE module (a combination of an RBMT system plus an SMT system), Senellart et al. (2010) concluded that the SPE system was superior to both the SMT and RBMT systems. Unlike large-scale evaluations which provide no detailed analysis as to the strengths and weaknesses of each participating system, fine-grained analysis could be found in these smaller-scale evaluations. Dugast et al. (2007) compared the output of an SPE system (Moses plus Systran) and an RBMT system (Systran)

---

<sup>39</sup> NIST 2008 Open MT Evaluation – Official Evaluation Results:  
[http://www.itl.nist.gov/iad/mig/tests/mt/2008/doc/mt08\\_official\\_results\\_v0.html](http://www.itl.nist.gov/iad/mig/tests/mt/2008/doc/mt08_official_results_v0.html) [last visited 2010-04-28]

and provided an in-depth report on the improvements and degradations of the SPE system.

Comparison can help consumers decide which system to use. An example is mentioned by TAUS<sup>40</sup> (2009) whereby Autodesk<sup>41</sup> decided to deploy Moses in production mode after their own experiment on Systran, Apertium (an open-source RBMT system) and their comparison of Systran against Moses. For researchers, comparison can help pinpoint a system's problems and devise methods for improvement. A comparison of the systems involved in this study will be presented in the last chapter.

## 2.3 Summary

In this chapter, a brief review of the main types of MT and representative examples of each type were presented. We first discussed the architecture of RBMT and detailed its translation process using Systran as an example. Systran's three internal translation processes were explained. As to the state-of-the-art SMT system, Moses was introduced. We especially focused on the explanation of the phrase-table in an SMT system as this is of significance to the present study. Their advantages and disadvantages have promoted the development of system combination. MT systems can be combined by way of using an SMT system to post-edit outputs of an RBMT system, or by adding the phrases from an SMT/RBMT into the dictionaries or phrase table of an RBMT/SMT.

The second part of this chapter reviewed MT evaluation including human and automatic evaluation as well as their application in many evaluation campaigns.

---

<sup>40</sup> Translation Automation User Society (TAUS): <http://www.translationautomation.com/> [last visited 2010-04-28]

<sup>41</sup> Autodesk's website: <http://usa.autodesk.com/> [last visited 2010-04-28]

The most widely used attributes in judging the quality of an MT output are adequacy and fluency. However, recently, ranking has become a more and more popular means in large scale evaluation programs. Error analysis is also widely used in order to reveal the errors of an MT system. In addition to focusing on sentence or document level evaluation, constituent-level evaluation is also of importance.

Automatic evaluation metrics are fast (vs. slow) and cheap (vs. expensive). They have become an important part in the cycle of the development of an MT system. There are string-based metrics such as BLEU, GTM and TER and linguistic-rich metrics. However, it is now widely recognised that no automatic metric can fully replace the role of human evaluation. Ideally, both evaluation methods should be used in order to fully assess the quality of an MT system. Correlations, including inter-evaluator, intra-evaluator and correlation between automatic and human evaluation, are usually examined and reported to measure the validity of research findings.

## **Chapter 3: Contextualising the Research Question**

### **- Translation of Prepositions**

Most MT systems nowadays can offer valuable help for information gisting (Krings 2001). However, this study is rooted in a localisation context in which the translation of technical documents is domain-specific and requires publishable quality. The role of an MT system in a localisation setting is to produce draft translations of documents, which will then be passed on to human post-editors to produce the final translations. Hence, the importance of post-editing cannot be overstated. Its importance can be justified not only by providing improved MT output, but also by the fact that post-editing can help improve the translation system. Post-editors can collect recurring errors and report them to the MT system developers or users, in some cases, with a suggestion on how to correct the system's dictionaries and linguistic components. Symantec, for example, compiles such reports with the help of internal professional translators.

One benefit of our industry-academia collaboration is that we could avail ourselves of this report from Symantec. In other words, we have access to the core problems of the MT system that severely affect the productivity of translators. Since these problems influence the timeline for product launch, resolving them is a top priority.

The sections of this chapter are arranged as follows. Section 3.1 begins with a brief review of the errors in MT output, followed by the sample error report from the internal translators at Symantec. Section 3.2 narrows down the research to one specific syntactic structure – prepositional phrases. A short introduction to the characteristics of English and Chinese prepositions will be provided and discussed.

An error typology of translation of prepositions will be established. Finally, based on the error typology, Section 3.3 states the research question of this study and Section 3.4 introduces several widely used pre- and post-processing approaches that may be useful in answering our research questions.

### 3.1 Errors in MT Output

Error classification is useful for both MT users and MT developers (Flanagan 1994). It is an efficient approach to evaluating the performance of an MT system in the sense that it can help pinpoint the problems of the engine and set a path for further research.

Not all MT errors are universal for all MT systems nor are they shared across all language pairs. Font Llitjós et al. (2005: 89) summarised an error typology for an RBMT system as follows (Figure 3.1).

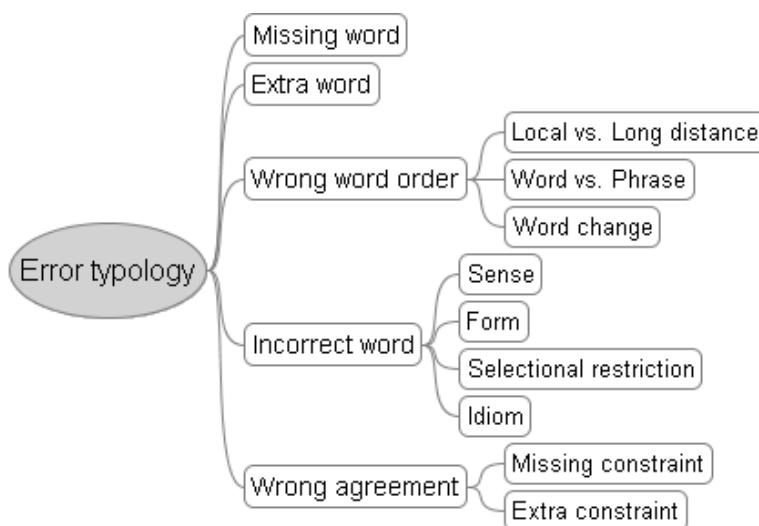


Figure 3.1: Error typology for an RBMT system

Based on this error typology, Vilar et al. (2006: 3) developed a fine-grained error typology for an SMT system (Figure 3.2).

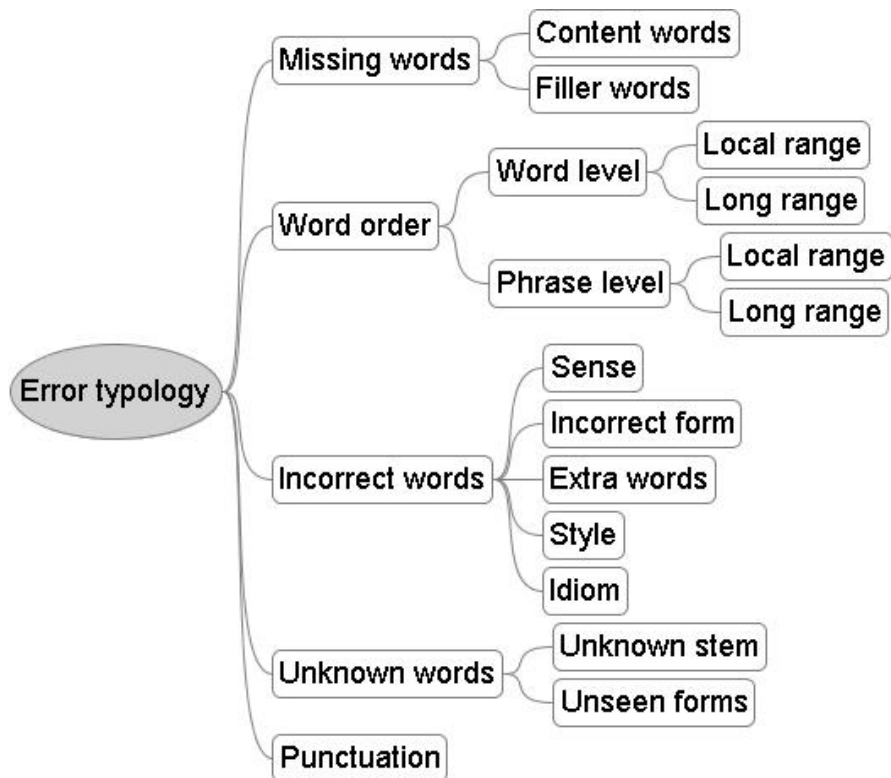


Figure 3.2: Error typology of an SMT system

In addition to general summarisation, there are also more detailed reports on errors. By observing each step of translation, Knowles (1978) found that one of the errors in the generation step is incorrect target generation of prepositions. While pointing out that unique category sets should be developed for different language pairs in order to reflect the error types that actually occur, Flanagan (1994) also showed that some errors are shared across languages. For example, the following errors were found in MT outputs which were translated from English into French and German: rearrangement error (sentence elements ordered incorrectly), preposition error (incorrect, absent or unneeded preposition), expression error (incorrect translation of multi-word expression), word selection error, etc.

To keep a record of the recurring errors of Systran, Symantec uses a tool called Etrack. It is a tracking system Symantec created in 2001 which was

originally used by users/developers of Symantec to report and monitor software bugs. Reports on translation using Systran can focus on dictionary problems or other linguistic problems. The research question of the current study originates from this error report. Table 3.1 is a summary of all the translation errors reported by the Chinese translators of Symantec. Four main categories are identified and the percent of each category among all the errors is reported on below.

Type	Example	Percent
Word/Term	Terms such as <i>mount</i> , <i>spring</i> and <i>slide</i> are translated incorrectly in their contexts.	17%
Clause	Translation of time clause <i>when...</i> sometimes it is generated in the wrong place.	33%
Preposition	Preposition <i>for</i> is often translated incorrectly. Translation of preposition <i>on</i> is often in an incorrect position in the translation.	33%
Others	Put space between English and Chinese characters in translated file.	17%

Table 3.1: Summary of errors reported by internal translators

It is worth noting that this report was extracted at the beginning of this project in 2008. Since then, many attempts have been made by both Systran and Symantec to tackle these problems in order to improve the quality of Chinese translation. A number of errors have already been solved, such as the first and the last error examples in Table 3.1. However, the two main challenges, translation of clauses (e.g. subordinate clauses) and translation of prepositions still remain. Two types of errors are associated with the translation of clauses and prepositions, i.e. incorrect lexical translation and incorrect word order.

These two challenges (clauses and prepositions) are by no means specific to the RBMT used (Systran) or the language pairs (English to Chinese) involved. As



mentioned at the beginning of this section, errors in the translation of prepositions are reported in many studies (Wu et al. 2006; Li et al. 2005; Flanagan 1994; Knowles 1978).

Differences between English and Chinese are the major reason for the above-mentioned problems. Take the translation of English attributive clauses into Chinese for example, unlike English, which puts attributive clauses after the nouns they modify, Chinese uses the attributive clauses directly before the nouns they modify. This structural difference raises non-trivial problems for an MT system. Translating an attributive clause from English into Chinese entails not only lexical substitution but a word order shift. Given the large number of subordinate clauses and relative clauses in English and their various translation equivalences in Chinese it is still hard to come up with the right rule to represent some structures (Arnold et al. 1994). Moreover, to modify the rules of an RBMT system requires linguistic resources and the overall quality of the translation is not always guaranteed to be better (Costa-Jussà et al. 2010).

English prepositions are another significant source of ambiguity. According to Saint-Dizier (2005), the English preposition is probably the most polysemic category and its linguistic realisations are extremely difficult to predict. For instance, there are more than 20 meanings of *with* according to the Longman Online English Dictionary, each of which may have a different Chinese translation.<sup>42</sup> Identifying the appropriate meaning of a preposition in a specific context is one of the hardest problems for MT (Saint-Dizier 2005).

Among the two main problems identified in the post-editors' report, the preposition is chosen as the core research question here. Let us define our research

---

<sup>42</sup> Longman Online English Dictionary: <http://www.ldoceonline.com/> [last visited 2010-06-27]

topic more clearly. In his study of translation of noun phrases/prepositional phrases (NP/PP) of an SMT system, Koehn (2003) described five levels of syntactic structures that an MT system has to take into account (Figure 3.3).

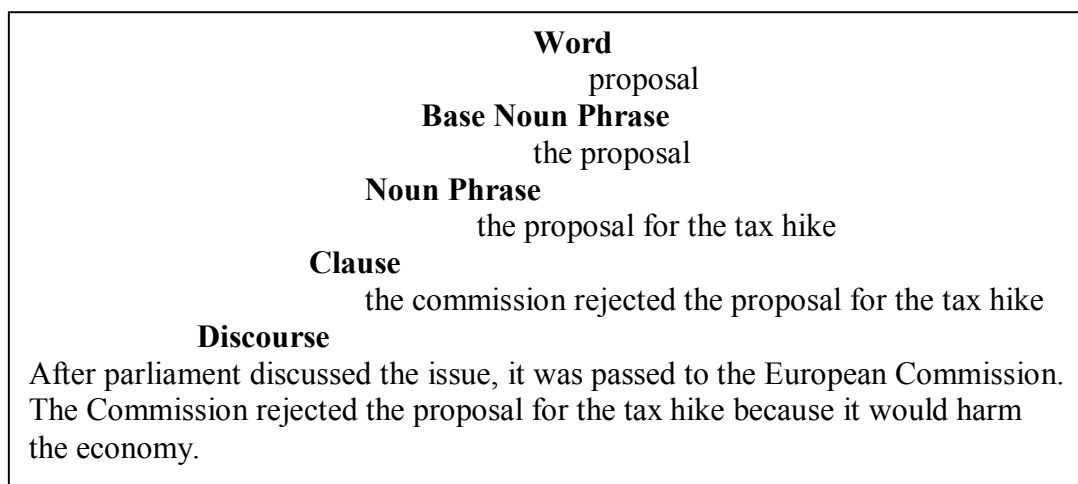


Figure 3.3: Five levels of syntactic structures (Koehn 2003: 2)

If we are to apply the above stratification to our study of prepositions, there are mainly four syntactic structures: a Preposition (P), such as *for*; a Prepositional phrase (PP), such as *for the clients*; a Clause, such as *the installation instructions for the clients are included in the package*; and Discourse, such as *A client bought a piece of software. The installation instructions for the client are included in the software package*. This study considers both the word level (P) and the phrase level (PP).

The most important reason of focusing on prepositions is that preposition is a closed word category (i.e. there are a fixed number of members) (Stott and Chapman 2001). Therefore, “a more exhaustive study of the linguistic properties” of the syntactic structures can be conducted (Koehn 2003: 2). In addition, more complex (or “computationally more expensive”) methods can be applied and monitored (Koehn 2003: 2).

However, it is important to bear in mind that the translation of prepositions intertwines with translation of clauses and other constituents of a sentence. In most cases, a preposition error can only be identified and tackled by taking the whole sentence into consideration. Therefore, we aim at proposing methods that can solve preposition problems in the context of the sentence rather than methods aimed at the translation of prepositions in isolation.

Before we focus on the preposition errors in the RBMT system, it is necessary to look at the general characteristics of English prepositions and Chinese prepositions. It is important to remind the readers that this study focuses on identifying the challenges that English prepositions pose to an RBMT system and how to improve the performance of the RBMT system in translating prepositions into Chinese. Hence, rather than illustrating the complexities of English to Chinese translation of prepositions in general, this study seeks to explore productive approaches to improve the output of the RBMT on the basis of a series of experiments. As such, the study is exploratory and experimental in nature although grounded in real-world contexts. The characteristics of English and Chinese prepositions will only be briefly introduced in this section so that readers understand the causes behind translation errors.

## **3.2 English and Chinese Prepositions**

### **3.2.1 English Prepositions**

Prepositions enjoy high frequency in English. There are simple prepositions such as *to*, *from*, *on*, *off*, and complex prepositions including *on top of*, *in front of*, *at the bottom of*. Prepositions indicate semantic roles encoding relational information (Hartrumpf et al. 2006). The relation expressed by a preposition is represented by

a preposition and its complement on the one side and another part of the sentence at the other side (Quirk et al. 1985). Therefore, a prepositional phrase is typically made up of a preposition plus “a noun phrase or a nominal *wh*-clause or a nominal *-ing* clause” (Quirk et al. 1985: 657).

English prepositions share some similarities with other word classes and constructions, such as particles, adverbs and, especially, subordinate conjunctions. The prepositions in English can be briefly defined in three ways, i.e. prepositions cannot have a complement that is a *that*-clause, or an infinitive clause, or a subjective case form of a personal pronoun (Quirk et al. 1985: 658-659).

There is a large volume of studies describing English prepositions linguistically, often from various angles. For example, Pullum and Huddleston (2002) investigated the syntactic functions of prepositions. Prepositions have been described from a cognitive perspective by Lakoff and Johnson (1980). A detailed semantic explanation can be found in the work of Saint-Dizier and Vazquez (2001), Saint-Dizier (2005) and Litkowski and Hargraves (2005). And, a pragmatic approach to prepositions has been examined by Fauconnier (1994).

A number of studies attempted to explain English prepositions in a way to assist natural language processing (NLP), one of which is The Preposition Project (TPP).<sup>43</sup> TPP attempts to provide a comprehensive characterisation of the meanings of English prepositions which would be suitable for NLP. Currently, 334 prepositions (mostly phrasal prepositions) are included with 673 senses identified. The semantic roles and the syntactic properties of the complements of

---

<sup>43</sup> TPP (The Preposition Project): <http://www.clres.com/prepositions.html> [last visited 2010-06-28]

each preposition are characterised in this project and reported by Litkowski and Hargraves (2005).

Although prepositions have been described by many studies, not all knowledge can be encoded into an MT system. One known problem with prepositions faced by any MT system is the PP attachment structure (Mamidi 2004). Possible PP attachment is denoted by a 4-tuple  $\langle V, N1, P, N2 \rangle$  where V denotes verbs; N1 denotes the object of the verb, usually preceding the preposition; P denotes a preposition and N2 another noun, usually following the preposition (Brill and Resnik 1994). There are several parsing options for a PP attachment structure. In NLP, syntactical parsing is a process of analysing the sentence into its grammatical structure according its grammar. It is of vital importance for almost every area of NLP, such as MT, questioning and answering (Q&A), etc. (Jurafsky and Martin 2009). Olteanu and Moldovan (2005) and Arnold (2003) showed that ambiguity of a sentence increases exponentially with the number of PP attachments in the sentence. For a structure containing one PP attachment, there are two parsing possibilities. For a chain of 2, 3, 4 and 5 PP attachments, there are 5, 14, 42 and 132 parsing options respectively. Let us look at a variant of the well-known example used to demonstrate the PP attachment problem: two parsing possibilities of one PP attachment in a sentence (example 3.1).

---

Example 3.1

**Source:** The police saw the man **with** a telescope.

**Parsing 1:** The police [saw [the man with a telescope]].

**Ref:** 警察 看见 了 那个 拿 望远镜 的 人 。 /pīnyīn: ná/

**Gloss:** Police saw LE (tense marker) the carrying telescope DE (modifier marker) man.

---

---

**Parsing 2:** The police [saw [the man] [with a telescope]].

**Ref:** 警察 用 望远镜 看见 了 那 个 人 。 /pīnyīn: yòng/

**Gloss:** Police use telescope saw LE (tense marker) the man.

---

For the interest of the readers who do not understand Chinese, we use Chinese *pīnyīn* (the Romanisation system of Chinese words indicating the way a word sounds) in addition to Chinese characters to clearly represent the translations of the highlighted English prepositions. In the examples below, different sound notation indicates different words with different meanings. The Chinese sentences are segmented (cf. Chapter 2) by the author in the way that each Chinese word is a translation equivalence of a source English word. Word-for-word glosses are also provided. However, as Chinese is not a morphologically rich language compared to English and tenses and modification relations in Chinese are indicated through function words instead of morphological changes, it will be difficult to back-translate these Chinese function words into English. Therefore, we spell out the sound of these words and comment on their functions in the brackets.

In example 3.1, there are two parsing possibilities of the same sentence. *Source* refers to the source English sentence and *Ref* refers to a TL translation of the source sentence. Both the English preposition and its Chinese translation are highlighted. The sound of the Chinese translation of the preposition is put at the end of the Chinese sentence within paired slashes //. In the glosses, each English word corresponds to one Chinese word, i.e. the first glossed English word corresponds to the first Chinese translation word and so on. Special strings such as “LE” or “DE” in the glosses correspond to the function words in the Chinese translation with their functions explained in brackets.

The source sentence in example 3.1 is ambiguous even for a human, not to mention an MT system. The PP attachment here conforms to the 4-tuple structure, i.e. V (*saw*), N1 (*the man*), P (*with*), N2 (*a telescope*). The PP (*with a telescope*) can either be parsed as an attributive of N1 (parsing 1 in example 3.1), or be attached to the verb V (parsing 2) as an adverbial. Two ways of parsing indicate two different understandings of the sentence expressed through different word order and translations of *with*.

In technical documents, sentences with PP attachment may cause ambiguity only for MT rather than for humans as illustrated in example 3.2.

---

Example 3.2

**Source:** Separate email addresses **with** commas.

**Parsing 1:** [Separate [email addresses]] [with commas].

**Ref:** 用 逗号 分开 邮件 地址 。 /pīnyīn: yòng/

**Gloss:** Use comma separate email address.

**Parsing 2:** [Separate] [email addresses with commas].

**Ref:** 分开 带 逗号 的 邮件 地址 。 /pīnyīn: dài/

**Gloss:** Separate containing comma DE (modifier marker) email address.

---

In this example, theoretically, *with commas* has two attachment options, one is to attach it to the noun *email addresses* and the other is to the verb *separate*. However, real world knowledge tells us that the second parsing (parsing 2) is incorrect as commas are usually not allowed in email addresses. Again, different meanings need different parses, which result in different word order and different corresponding translations.

The examples (3.1 and 3.2) show that if parsing is incorrect in the analysis step, the output of the translation is likely to be incorrect with different lexical

selection and problematic word order. Word order is one of the most important factors for determining the meaning of a sentence in Chinese.

Mamidi (2004) identified another two characteristics of English prepositions that pose challenges for an MT system, namely:

- (1) Semantically meaningful vs. semantically empty prepositions, i.e. deciding whether the preposition is part of a fixed phrase. For example (3.3a vs. 3.3b):

---

Example 3.3a

**Source:** Information about license keys is stored **on** the master server.

**Ref:** 关于 许可证 密钥 的 信息 存储 在 主 服务 上 。 /pīnyīn: zài...shàng/

**Gloss:** About license key DE (genitive marker) information store on master server on.

Example 3.3b

**Source:** He did this **on purpose**.

**Ref:** 他 故意 这样 做 。 /pīnyīn: gù yì/

**Gloss:** He on purpose this way does.

---

Preposition *on* in example 3.3a is semantically meaningful while *on* itself in example 3.3b is semantically empty and has to be bound with *purpose* to form a collocation (or an idiomatic prepositional phrase “*on purpose*”). These prepositional phrases are usually turned into adverbial constructions in Chinese which do not need a preposition. In technical documents, semantically empty prepositions (prepositions in collocations) do not cause serious problems for an MT system since the number of these phrases is small and translations of phrases like this (e.g. *such as, for example*) are usually not ambiguous. Another challenge is:



(2) Polysemous prepositions or various target language equivalences. For example (3.4a and 3.4b):

---

Example 3.4a

**Source:** To push the software **to** all clients.

**Ref:** 要 将 软件 介绍 给 客户 。 /pīnyīn: gěi/

**Gloss:** To JIANG (active voice marker) software push to client.

Example 3.4b

**Source:** They may be a threat **to** all clients and to their data.

**Ref:** 他们 对 客户 及 其 数据 来说 , 可能 是 威胁 。 /pīnyīn: duì/

**Gloss:** They to client and their data LAISHUO (complement words. Together with Dui, they mean “from the point of”), may be threat.

---

The highlighted preposition *to* in example 3.4 has different meanings. When translating into Chinese, the same *to* in the two sentences require different lexical selections. In addition, depending on its meaning, its realisation in Chinese also requires different word order with the former appearing after the verb and the latter before the verb.

These characteristics of English prepositions pose various challenges for an MT system. Failing to deal with any of the challenges may generate errors in the target output. Furthermore, the characteristics of Chinese prepositions can also add more challenges to translation from English into Chinese. The next section briefly introduces relevant characteristics of Chinese prepositions.

### 3.2.2 Chinese Prepositions

The Chinese preposition is not a closed word class, i.e. new members continue to appear. Another name for the Chinese preposition is “coverb” referring to a specific set of verbs in the Chinese language which are similar to English

prepositions (Yip and Rimmington 2004; Li and Thompson 1981). They are called coverbs because they have to be used in conjunction with other verbs in a sentence. In addition, most prepositions in Chinese are derived from verbs, and most of them can still function as verbs (Zhu 2004). For example, 在 /pīnyīn: zài/ can function both as a verb and a preposition (Yu 1994). In example 3.5 it is used as a verb in the first sentence but as a preposition in the second sentence.

---

Example 3.5

**Source:** 这本书在我这里。 /pīnyīn: zài/

**Ref:** The book is here.

**Source:** 在黑板上写字。 /pīnyīn: zài...shàng/

**Ref:** To write on the blackboard

---

The nature of Chinese prepositions means that in some situations Chinese prepositions should be translated into English verbs. Similarly, some English prepositions do not always need to be translated into Chinese prepositions.

Another special characteristic of Chinese prepositions is that some prepositions consist of a preposition character and a postposition character, which in general are called circumpositions (Liu 2002). In his comparative study of circumpositions in Chinese and other languages, Liu stated that:

框式介词,即在名词短语前后由前置词和后置词一起构成的介词结构...框式介词本质上是一种句法组合现象,而不是一种词汇现象 [Circumposition refers to a type of prepositional structure which consists of a preposition before a noun phrase and a postposition behind the noun phrase...it is a syntactic pattern rather than a lexical category (Liu 2002: 1)]

In English, circumposition is rare, but examples do exist such as *from that time on*.

In comparison, circumpositions are very common in Chinese (Liu 2002). Many

simple prepositions in English need to be translated into Chinese circumpositions. The difficulty lies in the fact that circumposition is not compulsory; instead, it is context-dependent. In the following example (example 3.6), the translations are equally correct with and without circumposition. From here on, we use *Systran* to refer to MT output from Systran.

---

Example 3.6

**Source:** He often practices playing guitar **in** the park.

**Systran:** 他 经常 在 公园 练习 弹 吉他 。 /pīnyīn: zài/

**Ref:** 他 经常 在 公园 里 练习 弹 吉他 。 /pīnyīn: zài...lǐ/

---

Both translations (*Systran* and *Ref*) in example 3.6 are correct translations of the source English sentence. The only difference between the MT output and the reference sentence in example 3.6 is their translation of the preposition *in*. From the highlighted words we can see that the reference sentence uses a circumposition while the MT output uses a single preposition. As both are correct translations, no glosses are provided.

However, for most instances expressing position, circumpositions should be used in Chinese; otherwise the translation will be extremely awkward, or at least difficult to understand. For example, the Systran output of the English sentence in example 3.7 is different from its reference translation in its lexical generation of the preposition *on*. In this example, without the circumposition, the translation is quite awkward sounding to a native Chinese speaker.

---

Example 3.7

**Source:** You can run the *rman* command from a command line **on** the client.

**Systran:** 您 能 在 客户端 从 命令 行 运行 *rman* 命令 。 /pīnyīn: zài/

---

---

**Gloss:** You can on client from command line run rman command.

**Ref:** 您 可以 在 客户端 上 通过 命令 行 运行 rman 命令 。 /pīnyīn:  
zài...shàng/

**Gloss:** You can on client on from command line run rman command.

---

As to when an English preposition should be (or should not be) translated into Chinese circumpositions, readers can refer to general studies of Chinese and English grammar and translation between English and Chinese such as Xu (2003).

In summary, translation from English into Chinese by an MT system is affected greatly by the characteristics of and the discrepancies between the English and Chinese languages. Having described the challenges an MT system faces in the task of translating English prepositions into Chinese, we will now move on to summarise the errors exhibited by the MT system used in this study. Based on the error classification, the research goal and research question are established.

### 3.3 Setting up a Preposition Error Typology

To find out how to improve the translation of prepositions, one has to know what the errors are. At the beginning of the chapter, we mentioned some prior work on error analysis. However, that work concentrates mainly on translation of texts or sentences instead of a specific syntactic constituent. The error report from the internal translators of Symantec is not preposition-specific either. The translators examined the translation of whole texts and recorded all the errors, one of the largest categories of which was the translation of prepositions.

Therefore, our first task is to set up an error typology for the translation of English prepositions into Chinese. Based on the work by Flanagan (1994) and

Font Llitjós et al. (2005), together with the short review of the characteristics of English and Chinese prepositions, we define the following four error categories:

- (1) Incorrect lexical selection: incorrect selection of translation of a preposition/PP. This includes cases where the translation of a preposition/PP has to be changed while requiring no word order change (example 3.8).

---

Example 3.8

**Source:** To add computers **to** the organizational unit.

**Systran:** 添加 计算机 **对** 组织 单位 。 /pīnyīn: duì/

**Gloss:** Add computer to organisational unit.

**Ref:** 将 计算机 添加 **到** 组织 单位 。 /pīnyīn: dào/

**Gloss:** JIANG (active voice marker) computer add to organisational unit.

---

- (2) Incomplete translation: in cases where, without a circumposition, the translation of prepositions is not complete (example 3.9).

---

Example 3.9

**Source:** **In** the Security Status dialog box, review the features that trigger a specific status.

**Systran:** 在 安全 状态 对话框 ， 请 查看 触发 具体 状态 的 特点 。  
/pīnyīn: zài/

**Gloss:** In Security Status dialog box ...

**Ref:** 在 安全 状态 对话框 **中** ， 请 查看 触发 某 状态 的 特征 有 哪  
些 。 /pīnyīn: zài...zhōng/

**Gloss:** In Security Status dialog box in ...

---

- (3) Incorrect position (mostly for prepositional phrases): the position of the translation of a preposition/PP has to be changed in the target sentence,

including a change at word level, i.e. the position of the translation of a single preposition has to be altered in the target sentence; and a change at phrase level is required where the translation of an entire phrase has to be moved. In example 3.10, the cause of this error is the ambiguous PP attachment “*about infected computers*”.

---

#### Example 3.10

**Source:** Add a warning to email messages **about infected computers**.

**Systran:** 添加 警告 对 **关于 受感染的 计算机** 的 电子 邮件 。 /pīnyīn: guān yú/

**Gloss:** Add warning to about affected computer DE (genitive marker) email message.

**Ref:** 向 电子 邮件 中 添加 **关于 受感染 计算机** 的 警告 。 /pīnyīn: guān yú/

**Gloss:** To email message in add about affected computer DE (modifier marker) warning.

---

- (4) Translation missing: no correct translation of a preposition/PP is found. In the example below (Example 3.11), the MT failed to produce any translation of the highlighted preposition. The reason for the error in this example again originates from the incorrect attachment of the PP “*for the following Microsoft Exchange server versions*”. Hence, in this example, error 3 also exists.

---

#### Example 3.11

**Source:** The client software creates file and folder scan exclusions **for** the following Microsoft Exchange server versions.

**Systran:** 客户端 软件 创建 文件 和 文件夹 以下 Microsoft Exchange Server 版本 的 扫描 排除 。

**Gloss:** Client software creates file and folder following Microsoft Exchange

---

---

Server version DE (genitive marker) scan exclusion.

**Ref:** 客户端 软件 为 下列 Microsoft Exchange Server 版本 创建 文件 和 文件夹 的 扫描 排除 项 。 /pīnyīn: wéi/

**Gloss:** Client software for following Microsoft Exchange Server version create file and folder DE (genitive marker) scan exclusion item.

---

Readers may have noticed from the glosses that errors do not appear alone, but to the contrary, there are both lexical selection and word order errors in most of the examples we discussed. Lexical selection errors exist because English prepositions are polysemous and their corresponding equivalences are variable. Some English prepositions need to be translated into Chinese circumpositions instead of single prepositions (see examples 3.6; 3.7; 3.9).

One cause of word order error is the English PP attachment structure we have just discussed. Incorrect analysis of the structure by an MT system may generate incorrect target word order due to the grammatical differences between English and Chinese. For example, Wu et al. (2006: 601) pointed out that English prepositional phrases functioning as post-posed modifiers of nouns usually correspond to Chinese pre-posed attributives. When these phrases are transferred into Chinese equivalences, apart from the change of word order, one important feature is that between these phrases and the head noun, a structural particle such as DE (see example 3.1; 3.10) should be added. Li and Thompson (1981: 409) summarised that in Chinese, adverbial PPs usually occur before a verb and complement PPs occur after a verb while in English both types of PPs usually occur after verbs. There have been numerous studies trying to clarify various translation possibilities between English prepositions and Chinese equivalences both in the domain of MT and in the more general domain (Wu et al. 2006; Li and

Thompson 1981). However, due to the complexity inherent in languages, there are still cases where researchers find it difficult to describe the nature of the problem precisely. Therefore, no RBMT system at the moment is equipped with all the rules needed to transfer all structures.

To sum up, due to the differences between English and Chinese as well as the unique characteristics of prepositions in the two languages, many English prepositions are generated incorrectly into Chinese by the RBMT system. Our ultimate research goal is to improve the translation of prepositions and, consequently, improve the overall translation quality by reducing these errors. The next section puts forward the research questions.

### **3.4 Research Questions**

As mentioned, this study does not intend to establish a classification of the various translation equivalences between English and Chinese prepositions. What we are interested in are the errors produced by the RBMT system and how to reduce the errors. Hence, the main research question is how to improve the Machine Translation of English prepositions into Chinese by an RBMT system operating in the IT domain. This question can be broken down into the following sub-questions and we anticipate that together their answers will contribute to answering the main research question:

- *Question 1: Which prepositions are translated incorrectly?*
- *Question 2: Which errors occur most frequently in our selected corpus?*
- *Question 3: What type of errors are associated with each preposition?*
- *Question 4: What existing solutions are suitable for tackling the most common errors?*



- *Question 5: What are the possible effective solutions that have not yet been tested?*

We will deal with the first three sub-questions in Chapter 5 where a detailed evaluation of the errors associated with prepositions is conducted and reported. Chapters 4, 6, 7 and 8 answer the last two research questions. For now, let us briefly review some of the general approaches proposed to improve the performance of an RBMT system. This will provide some additional context before we discuss our methodology in detail in the next chapter.

### **3.5 Further Context**

To date, numerous studies have been undertaken with the aim of obtaining a better translation of prepositions. Research of this kind, however, investigates the problems caused by prepositions mainly from the point of view of developers, i.e. by modifying or controlling the system architecture, such as proposing more/new transfer rules (cf. Chapter 2). For example, Hartrumpf (1999) combined interpretational rules and statistical methods to improve PP attachment disambiguation and preposition interpretation. Gustavii (2005), in his experiment, showed that using transformation-based learning to induce rules from aligned bilingual corpora could help select the appropriate target language preposition.

However, as a general user of the RBMT system, the author (and Symantec) were not in a position to add new language transfer rules. Hence, the black-box approach was the only option. The above-mentioned approaches cannot be applied in this study because, unlike system developers, we have no access to the internal translation process. Researchers in similar scenarios pursue improvement either by pre-processing (i.e. work on the source text before inputting it to an MT

system) or post-processing (i.e. work on the target output after the source text is machine-translated).

With regard to RBMT systems, available customisation, pre-processing and post-processing approaches include dictionary customisation, controlled authoring and post-editing. The localisation department within Symantec makes full use of various pre- and post-processing approaches to achieve high quality output. As mentioned in Chapter 2, Systran provides its users with the option to build their domain-specific user dictionaries (UD). Symantec, therefore, has created its own domain-specific dictionaries for all the language pairs localised by the company, each of which contains several hundred to thousands of entries compiled by in-house linguists. According to Systran's user manual, its IntuitiveCoding technology allows users to include additional linguistic information, such as a word's collocational preposition. Figure 3.4 shows an excerpt from the Help file of Systran about encoding entries into a dictionary.

English	French
accountable (prep: for)	responsable (prep: de)
request (prep: for)	demande (prep: de)
to dream (prep: of)	rêver (prep: de)

Figure 3.4: Systran's (v.6) instruction sample for encoding dictionary entries

Authoring by controlled language (CL) is also a widely-used pre-processing method. CL is defined as “an explicitly defined restriction of a natural language that specifies constraints on lexicon, grammar, and style” (Huijsen 1998: 2). The mechanism of CL is to minimise ambiguities from the source instead of correcting errors after translation. O'Brien (2006) provided empirical evidence that controlling the input to an MT system could lead to faster post-editing speed

indicating improvement in MT output. This finding is supported by Roturier (2006) who showed that CL rules could improve translatability and readability of the output of an RBMT system. In order to ensure that documents are written in a way that conforms to the rules specified, CL checkers have been developed. These checkers can flag sentences violating predefined rules, and thus can help writers focus on only ambiguous structures. An example of such a checker is acrolinx's acrolinx IQ<sup>TM</sup>.<sup>44</sup> Like Systran, acrolinx IQ contains some general authoring rules while at the same time allowing its users to compose rules to meet their own needs. To give a simple example of how it works, suppose we want to flag all prepositions in a text, we can specify such a rule in acrolinx IQ, run the sample we want to check through acrolinx IQ, and all the prepositions can be highlighted in the output file. We applied this method for our study (see Section 4.3.2) to extract our preposition corpus.

There are both benefits and drawbacks of UD and CL. On the one hand, both can improve the output of an MT system. On the other hand, their implementation requires large numbers of human resources, which in turn implies both time and cost. Moreover, UDs and CL rules are considered to be confidential assets by the companies who used them. Hence, little work on user dictionaries has been reported and only a few CL rules on prepositions can be found in the literature (see Section 4.5.2).

Besides manipulating the source texts, post-processing the output is also quite important. To get output of publishable quality in a localisation context, human post-editing is today generally considered as a necessary step. However, as Allen and Hogan (2000) point out, MT errors are likely to recur throughout or across

---

<sup>44</sup> acrolinx: [http://www.acrolinx.com/why\\_acrolinx\\_iq\\_en.html](http://www.acrolinx.com/why_acrolinx_iq_en.html) [last visited 2010-06-29]

documents. Therefore post-editors are often dispirited by the need to make the same correction over and over again (Isabelle et al. 2007: 255). In order to ease the burden placed on human post-editors, several attempts have been made to reduce the work of post-editors by automatically correcting some recurrent errors. Semi-automatic search and replace (S&R) using regular expressions (RE) is one of them. Regular expressions refers to a special language for searching strings (including letters, numbers, spaces, tabs and punctuation) in a text (Jurafsky and Martin 2009). The advantage of using REs is that once an error pattern is found and defined, then all errors matching this pattern can be replaced automatically (Guzmán 2008). However, the drawback is also obvious. To find a match, a string has to be precisely defined. This can become very complicated even for a simple task. A good example is illustrated by Jurafsky and Martin (2009). Suppose we want to find the English article *the* and replace it with something else, then we have to take all the following situations into consideration:

- (1) First, there might be many variants of *the* such as *The*, *the*, *THE*, etc.;
- (2) Second, there might be false matches such as *other*, *theology*, etc. where *the* is embedded;
- (3) Third, some special contexts have numbers or underlines, such as *the\_*, or *the25*.

Another automatic post-editing idea first put forward by Allen and Hogan (2000) is the development of automatic post-editing (APE) that would automatically repair mistakes in raw MT output by utilising the information on the changes that were made during the post-editing process from “parallel tri-text” (source texts, MT output, post-edited texts) (ibid: 62). Elming (2006) presented the results of the use of an APE module to correct the output of an RBMT system

and it was noted that translation quality increased noticeably in terms of BLEU scores. The advent of SMT opened the door to the possibilities of combining two different MT systems to benefit from the advantages of both. Knight and Chander (1994) proposed to use SMT techniques to learn the mapping between a large corpus of “pre-edited” (1994: 779) texts with aligned corresponding post-edited texts. Simard et al. (2007a; 2007b) tested and extended this proposal by using an SMT system to post-edit the output of an RBMT system. As discussed in Chapter 2, this kind of module is now often referred to as an SPE module.

As Symantec employed all the above-mentioned methods in their production cycle, it was necessary to examine their benefits and drawbacks in order to answer our fourth and fifth sub-questions. The design of a pilot test as well as the whole research methodology is introduced in Chapter 4.

### **3.6 Summary**

This chapter defines the research questions and the research goal of the current study. Both Chinese translators’ reports and prior research work concur that translation of prepositions is one of the major challenges faced by the RBMT system. In general, the translation errors for prepositions from English into Chinese can be categorised into incorrect lexical selection and incorrect word order. One main cause of errors is the structural divergence between English (the source language in the context of this study) and Chinese (our target language), especially in terms of the word order of prepositional phrases. A second cause of errors is the polysemous nature of English prepositions and the special circumpositions in Chinese. Based on this general categorisation, we set up a fine-grained error typology of the RBMT system’s output of prepositions from English technical documents into Chinese. This error typology opens up a path for

our further research. We aim to explore which errors are the most frequent, which prepositions are most problematic, and how to improve the translation of prepositions. There are already pre- and post-processing approaches in the literature to improve the output of an RBMT system. These approaches are widely employed by industry and studied by researchers. However, the effectiveness of these approaches on the translation of prepositions has been less examined, especially from English into Chinese. This study further tests the usefulness of those pre- and post-processing approaches and proposes several new approaches.

## Chapter 4: Methodology

As Lee and Renzetti (1993) pointed out “the ‘what’ to investigate, must come prior to the decision of ‘how’ to go about doing one’s research” (p.27). Having decided *what* to study and outlined the research questions in the previous chapter, this chapter will elaborate on the procedures we intend to follow to answer the research questions.

Section 4.1 discusses briefly the settings of the main MT systems used in this study: Systran and Moses. The core objective of this study is to improve the performance of the RBMT system through the help of the SMT system.

Section 4.2 provides information about the specific form of human evaluation and the automatic metrics adopted for the purpose of this study. In order to check the quality of translations, extensive evaluation or comparison is required. MT evaluation can be done in two ways as introduced in Chapter 2, namely, human and automatic evaluation. Ideally, both evaluation methods should be used and qualitative and quantitative analyses should be conducted to fully assess the effects of the approach taken.

The process of corpus compilation is described in detail and particular in Section 4.3. This is followed by a discussion of the other principles of research design, including internal and external validities in Section 4.4. Finally, Section 4.5 explains how several pilot tests were conducted and reports on their impact on the subsequent research design.

## 4.1 MT Systems Used

In this study, two MT systems are employed. The Baseline MT system is Systran (version 6.0) used by Symantec. As mentioned, Symantec customised this MT system with its own domain-specific user dictionaries (UDs). Entries in these user dictionaries include either general terms with specific meanings pertaining to the IT domain or unique term that only appears in Symantec's documents. One can choose which dictionary or whether or not to use these dictionaries. The benefits of domain-specific UD's have been reported by Arnold et al. (1994) and Dugast et al. (2009). Our pilot test (see Section 4.5.1) comparing two translations of the same sample shows that without the UD's, term is often translated incorrectly. Since the corpus we use in this study is from Symantec, it is logical to use the Symantec user dictionaries in order to ensure correct translation of term. Additionally, the default setting of Systran employed by Symantec is with both the suitable general Systran dictionaries and the Symantec UD's. Therefore, the output from Systran with this setting is called the **Baseline** output in the study.

The second main MT system is Moses which is an open-source SMT toolkit as well as a stand-alone SMT system. It is currently the most widely used system by researchers (Koehn et al. 2007). This study used the Moses toolkit installed in Symantec following the tutorial on its website. Note that the version of Moses employed is the simplest version which deals with plain texts without any extra linguistic knowledge. Recently, more complex versions which require rich linguistic information such as tagging or parsing have been created. Since most approaches we propose in this study are general techniques, these more complex versions of Moses can also be applied to these approaches.



## 4.2 Evaluation – Measurement Validity and Reliability

Measurement, according to Frey et al., refers to “the process of determining the existence, characteristics, size and/or quantity of some variable through systematic recording and organization of observation” (1991: 100). Developing valid measurement is a primary concern for researchers, which indicates that researchers are indeed measuring the concepts they intended to measure and the variable is measured in a consistent and stable manner (Frey et al. 1991).

Measuring the effect of an approach on the translation of prepositions is one of the core research objectives of the study. To choose the appropriate human and automatic evaluation, specific problems have to be taken into consideration. To obtain an overview of the translation quality of prepositions of the RBMT system requires human examination so that questions like *which error is the most frequent* can be answered. Evaluation of the errors in translated prepositions is the first step taken in this study before we apply any approaches to improve the Baseline translation. The details of this evaluation are reported in Chapter 5.

As for measuring the effects of an approach, translations can be compared and evaluated both by humans and automatic evaluation metrics.

Automatic evaluation metrics can report in a quantitative way the scores of the overall translations which can reflect whether or not there is a difference between two translations. For this study, we selected three of the most widely used metrics, namely, BLEU, GTM and TER. Several factors influenced this decision: they are widely used in the area of MT evaluation; they are able to evaluate Chinese output; they are reported to correlate with human evaluation to some extent; they are straightforward to use; no additional large datasets of linguistic information are needed. One particular reason for the choice of GTM

was that it is the default evaluation metric embedded in SymEval (an evaluation software program used in Symantec) (Roturier 2009). In Section 2.2.2.2 where GTM was previously explained, we mentioned that the weight of GTM can be changed. The higher the weight, the more penalties on the word order difference between an MT output and its reference translation. The most commonly used weight is the default setting ( $e=1$ ) which applies no penalty to word order differences. Turian et al. (2003) concluded from their evaluation of Chinese output that ( $e=1$ ) correlated better with human evaluation than GTM ( $e=2$ ). Another common weight of GTM is  $e=1.2$  which is used in some evaluation campaigns (Callison-Burch et al. 2007). In addition, GTM ( $e=1.2$ ) is also internally employed by Symantec (Roturier 2009). Therefore, in this study, both the default ( $e=1$ ) and ( $e=1.2$ ) are reported throughout all experiments. Note that only automatic evaluation scores of a system and of each sentence were reported and examined. We did not extract isolated translation of prepositions to be scored as most automatic metrics are designed to work on text or sentence level instead of on short syntactic constituents.

Qualitative comparison of the differences between two translations requires human evaluation, particularly for preposition evaluation. Although the focus of the study is to improve the translation of prepositions, it is not desirable to obtain better translations of prepositions at the expense of lowering overall translation quality. Therefore, besides preposition evaluation, sentence level evaluation is also indispensable. Scoring an output at sentence level according to its adequacy and fluency is a pervasive evaluation approach (Flanagan 2009; Callison-Burch et al. 2007; LDC 2005; Hovy et al. 2002). However, more recent work has revealed that ranking is more intuitive, reliable and evaluator-friendly (Duh 2008; Vilar et

al. 2007). This type of evaluation is also found to be widely employed in some MT evaluation campaigns (Callison-Burch et al. 2009, 2008, 2007) by asking human judges to only rank the candidate translations from best to worst. For the purpose of this study, ranking at both preposition and sentence levels are conducted. The results are complemented by a detailed qualitative analysis of the outputs by the author.

There is no ideal profile described in the literature today as to the best evaluators. However, using professional translators who are familiar with the technical documents of this study would increase consistency and validity (Aranberri 2009). As to the adequate number of evaluators, many researchers pointed out that at least three or four evaluators should be used (Arnold et al. 1994; Carroll 1966). A minimum of four evaluators were employed in this study based on the above-mentioned information. Another consideration is the constraints of the research budget as the evaluation will become more costly with more evaluators involved.

The reliability of the results of automatic and human evaluation also needs to be examined. The reliability of human evaluation can be reported by the inter-evaluator and/or intra-evaluator correlation (such as Kappa scores) introduced in Section 2.2.3. The reliability of automatic evaluation can be verified by examining their correlation with human evaluation. Using both of them in a study can make use of the advantages of both and increase the overall validity of the results.

Frey et al. (1991) pointed out that “a researcher who intends to use a technique must make sure it has been validated at some point by its originators and has been used previously in research” (p.122). The measurement technique

used in this study meets this requirement as both the selected automatic metrics and the form of human evaluation in this study have been and remain widely used in the field of machine translation.

### **4.3 Corpus Design**

A corpus is a collection of material (text or speech) that is put together based on some criteria and serves the purpose of extracting information and/or for testing hypotheses (Megerdooonian 2003). In order to make meaningful interpretation of the results, special attention has to be paid to data preparation (Hatch and Farhady 1982). The principles of corpus design and compilation reported by Bowker and Pearson (2002) and Kennedy (1998) have been drawn upon throughout the study.

#### **4.3.1 Major Principles**

The following issues should be considered in designing a corpus: is the corpus general or specific; is it a static or dynamic corpus; what is the size of the corpus; how representative is the corpus; etc. (Kennedy 1998).

A corpus that is designed with particular research projects in mind is usually called a specialised corpus (Kennedy 1998). Since this study focuses on MT of technical documents, the corpus employed is a domain-specific (or specialised) corpus. One problem associated with a specialised corpus is that it may only provide a distorted view and may not be suitable for general purposes (Megerdooonian 2003). While this may be true, one has to bear in mind that the decision as to what corpus to use is determined by the research purpose, research questions and resources available. A specialised corpus is said to be advantageous in that various ambiguities are minimised and more controllable if the focus is on domain-specific material (Farghaly 2003).

Another issue related to corpus design is whether a corpus is static or dynamic (Kennedy 1998). A static corpus attempts to provide a “snapshot” of a language or a text type at a particular time, and usually no more text is included into it once it is built. A dynamic corpus, on the other hand, continues growing or updating over time. Although technical documents are subject to updates regularly, the problem of accurately translating prepositions via the RBMT system still remain. Moreover, being a closed word class in English means that few new prepositions appear over time (Stott and Chapman 2001). Therefore, the frequency or distribution of prepositions largely remains unchanged in domain-specific corpora. The discrepancy between the corpus we study and any updated version of the corpus is not likely to have a major impact on the findings concluded using this corpus.

Another concern is the representativeness of the corpus. The documents provided by Symantec include installation guides, user manuals and maintenance guides. An important question is whether or not our corpus represents, in any way, general language texts. The corpora were composed by technical writers who are native English speakers and are professional technical writers, thus a reasonable expectation is that prepositions are used in a more or less standard and native way and in a way that is appropriate to technical documentation in general. Our analysis of the frequency of specific prepositions in our specialised corpus and a comparison with the occurrence of prepositions in more general English-language corpora revealed interesting similarities (see Section 4.3.2 below). Hence, the prepositions studied are representative of the general use of prepositions in many types of corpora.

An essential element which influences the representativeness of a corpus is the corpus size (Kennedy 1998). Meyer pointed out that “to determine how long a corpus should be, it is first of all important to compare the resources that will be available to create it” (2002: 32). Bowker and Pearson (2002: 48) claimed that in studies related to Language for Special Purposes (LSP), corpora ranging between ten thousand to several hundreds of thousands of words had proven “exceptionally useful”. From his experiment, Biber (1990; 1993) found that valid and reliable information about the distribution of prepositions could be extracted from a sample of 1000 words. Based on this, Meyer (2002: 39) summarised that “if one studied the distribution of prepositions in the first thousand words of a newspaper article totalling 10,000 words, studying the distribution of prepositions in the entire article would not yield different distributions”.

The first corpus that was made available to the author was a corpus written conforming to Symantec’s in-house authoring rules which means the ambiguities in the source have been minimised. Since there is no rule regulating the use of prepositions, the distribution of prepositions was not affected by the controlled rules. Reference translations for these texts already existed; hence, we did not have to produce standard translations. The English corpus contains 204,412 words which falls within the range specified by Bowker and Pearson (2002). As we are interested in the prepositions in the sentences, from this corpus a preposition corpus (i.e. a corpus composed of sentences with at least one preposition) was extracted. It is described in the next section.

### 4.3.2 Preposition Corpus and Sample Extraction

To extract parallel sentences with prepositions, the in-house controlled language (CL) checker acrolinx IQ (cf. Chapter 2) was programmed to flag and output sentences with prepositions. Simply put, this in-house CL checker is based on pattern-matching. To analyse a text, the checker first assigns basic linguistic information, such as POS information, to each word. Next, it runs all rules for each sentence. If a pattern specified in a rule is found in a sentence, then the violated section of the sentence will be flagged in the result. Users of acrolinx IQ can create their own rules for a given special purpose. Some of Symantec’s rules include: “avoid use of passive voice”; “the length of a sentence should not exceed 25 words”. In this study, we use the checker to extract sentences with prepositions instead of extracting sentences violating certain rules. Therefore, we define a special “rule” as follows (Figure 4.1):

<pre>#ERROR Find_Prep  @prep      ::= [POS "^ IN"]  TRIGGER (80) == @prep</pre>
---

Figure 4.1: A sample rule defined in acrolinx IQ

The first line specifies the name of the rule. The second line declares the object, namely, everything with the POS of *IN*. The POS tag used by acrolinx IQ is the Penn Treebank tag set where the preposition is marked by *IN*.<sup>45</sup> The main rule specifies how to trigger a rule and what action to take if a rule is triggered. The first part of the rule contains a rule header, which specifies the type of rule (note that there are also rules other than *TRIGGER* rules) along with a confidence score

---

<sup>45</sup> Penn Treebank tag set can be found on: <http://www.cis.upenn.edu/~treebank/> [last visited 2010-08-15]

(the probability that an error can be confirmed). The second part of the rule is a pattern matching algorithm which tries to match the object defined in the second line. Once the object is matched, this rule will be triggered. An example of the output file is presented in Figure 4.2. Prepositions in each sentence are highlighted in red.

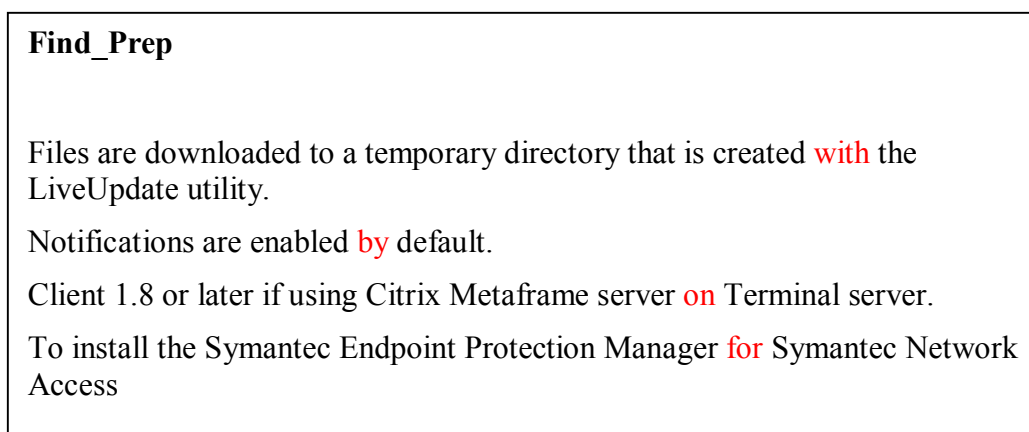


Figure 4.2: A screenshot of the output file of acrolinx IQ

What we just described is only the simplest scenario. In fact, *IN* does not only indicate prepositions but also subordinating conjunction. The preposition *to* is not represented by the *IN* tag but has its unique POS tag *TO*. Moreover, no tagger can achieve 100% precision at the moment which means that there are false triggers in the results. To ensure high precision, we carefully tested the extraction rule during several passes and accompanied it with a list of exceptions. The final preposition corpus contains 176,046 words. The number of times that each preposition occurs in the corpus is recorded and reported in Table 4.1.



Preposition	Frequency	Preposition	Frequency	Preposition	Frequency
in	1859	between	58	along	7
of	1782	within	43	beneath	4
for	1597	over	36	out	4
to	1326	without	36	throughout	4
on	1209	against	28	down	3
from	688	like	16	outside	3
with	541	below	14	per	3
by	393	beside	13	upon	3
about	344	up	12	versus	3
as	236	across	10	unlike	2
under	224	until	10	behind	1
at	167	above	8	near	1
during	79	except	8	toward	1
after	78	onto	7	towards	1
through	77	off	6	via	1
into	58	inside	5		
before	51	because	4		

Table 4.1: Frequency of each preposition in the corpus

It can be seen from Table 4.1 that the top ten most frequent prepositions make up the majority of all the prepositions in the corpus. In fact, 90% of all prepositions are represented by the top ten prepositions with only less than 7% occupied by the remaining 39 prepositions. These top ten prepositions are isolated in Table 4.2 with their relative frequency among the ten prepositions.

Preposition	in	of	for	to	on	from	with	by	about	as
Percent	19%	18%	16%	13%	12%	7%	5%	4%	3%	2%

Table 4.2: The top ten frequent prepositions and their relative frequencies

The distribution of the prepositions listed above was compared with that of other types of corpora to check if it is representative of other corpora. Table 4.3 below reports respectively the first ten most frequent prepositions in some other general corpora. The FrameNet (FN) corpus contains the 100-million-word British National Corpus, which itself contains texts of various genres (editorials,

textbooks, advertisements, novels and sermons) (Olteanu and Moldovan 2005).<sup>46</sup> Treebank 2 (TB2) consists of financial speeches and Wall Street Journal newspaper articles (Olteanu and Moldovan 2005).<sup>47</sup> The COBUILD (Collins Birmingham University International Language Database) is a 16-million-word corpus containing contracts, letters and theatre programs (Jurafsky and Martin 2009).<sup>48</sup> The LOB (Lancaster-Oslo-Bergen) corpus is a million-word collection of English texts about religion, trades and journalistic texts (Mindt and Weber 1989).<sup>49</sup>

Frequency Ranking	1	2	3	4	5	6	7	8	9	10
FN	of	to	in	for	on	with	from	<b>at</b>	as	by
TB2	of	in	to	for	from	on	with	by	<b>at</b>	as
COBUILD	of	in	for	to	with	on	<b>at</b>	by	from	about
LOB	of	in	to	for	with	on	by	<b>at</b>	from	as

Table 4.3: Frequencies of the top ten frequent prepositions in other corpora

A comparison of Table 4.2 and 4.3 reveals that most prepositions in our corpus are shared by corpora of other genres with only one exception (preposition *about*). The finding verifies that the distribution of prepositions in our corpus is representative of the distribution of prepositions in various types of texts.

In the end, the top ten prepositions were selected in this study since over 90% of all prepositions in our corpus were distributed among these ten prepositions. In addition, the fact that they are also in common with other corpora suggests that the results can be generalised to other studies. Finally, reducing the number of prepositions makes the error analysis and further qualitative analysis more controllable.

<sup>46</sup> FrameNet: <http://framenet.icsi.berkeley.edu/> [last visited 2010-06-29]

<sup>47</sup> Introduction to Treebank2: <http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T7> [last visited 2010-06-29]

<sup>48</sup> CELEX: [http://www ldc.upenn.edu/Catalog/readme\\_files/celex.readme.html](http://www ldc.upenn.edu/Catalog/readme_files/celex.readme.html) [last visited 2010-06-29]

<sup>49</sup> LOB: <http://khnt.hit.uib.no/icame/manuals/lobman/> [last visited 2010-06-29]

To machine translate all sentences with the ten prepositions and then obtain human evaluation for them all was considered impractical in view of time and available resources. Instead, based on the claim of Biber (1990, 1993) that 1000 word excerpts are enough to provide valid and reliable information on recurrent linguistic items, a random sample with 1000 prepositions was extracted from the preposition corpus, leaving the rest as the training data for the SMT system in the following chapters.

Sampling is a process of selecting individuals (or specific items) to take part in the research (Frey et al. 1991). An appropriate sample has to represent the population where the sample comes from. One of the practical methods suggested by Frey et al. (1991), called “purposive sampling”, is used here. First of all, sentences with the top ten prepositions were extracted from the corpus. Then, based on the relative frequency of each preposition listed in Table 4.2, sample sentences were extracted randomly. For example, *of* occurs in 18% of the sentences with the top ten prepositions. Therefore, we randomly selected 180 sentences with *of* ( $1000 \times 18\%$ ), and so on. Since some sentences have more than one preposition, the same sentence may be selected repeatedly. The author manually checked the extracted sentences in order to make sure that no preposition in the same part of a sentence was selected twice. In the end, 944 English sentences with 1000 preposition instances were extracted as the test sample.

## 4.4 Other Research Design Issues

Frey et al. (1991) mentioned that researchers might influence the judgments of participants by their unintentional expectancy. For example, in our human evaluation process, the results will be invalid if evaluators tend to give the expected answer due to an unintentional indication made by the researchers. To eliminate this threat, Frey et al. (1991) suggested removing researchers from the actual study or exposing the participants to exactly the same research environment. To make this research procedure as consistent and valid as possible, professional translators were employed by Symantec. They worked in their daily working environment without any face-to-face communication with the author. A third party sent the instruction and evaluation sheets to the evaluators and sent the results back to the author by email. All evaluators received the same instruction and the same set of data. This set-up demanded that instructions be extremely clear, unbiased and easy to follow for the human evaluators. In order to eliminate unnecessary misunderstandings or miscommunications, the author made her phone numbers and email addresses available to the evaluators in case communication was needed. However, none contacted the author through this channel and all the evaluations were conducted successfully. Once the data were generated, they were analysed in a systematic way.

Ecological validity refers to the research procedures which must reflect real-life situations in order to generalise the findings to other situations (Frey et al. 1991). We have established earlier that this research is rooted in the localisation production cycle of Symantec, which can be considered as a typical cycle among large volume localisation companies. In addition, as mentioned above, the RBMT system used is widely used by other companies, and documents such as user

guides and/or installation manuals are widely machine translated. Besides, the top ten prepositions investigated in this study are commonly distributed in other documents as well, as we have shown. Therefore, it can be argued that our findings can be generalised to other situations.

As the evaluation requires human participants, ethics approval was obtained from the school research committee. The evaluators were informed that their involvement in the study was voluntary, and that the data collated will be used only by the researcher and would not be given to anybody else.

## **4.5 Pilot Tests**

It has been pointed out that users often engage in pre- and post-processing to improve the translation of an RBMT system. Some of the pre- and post-processing methods include dictionary customisation, controlled authoring, etc (see Section 3.5 in Chapter 3). For our study, several exploratory pilot tests were carried out to check the effectiveness of these methods in improving the translation of prepositions. The objective of the pilot tests was to identify methods that might have the greatest potential for improving the translation of prepositions and to pursue those methods above others.

One pre-processing approach tested in the pilot phase was CL authoring rules. And two post-processing approaches involved were automatic search and replace (S&R) through regular expressions and statistical post-editing (SPE) through an SMT system. In addition, differences that domain-specific UD's can make to the translations were also examined. These are well-established methods that are also employed by Symantec currently to improve the output of the RBMT system.

As this is just an exploratory pilot test, only one automatic evaluation metric GTM ( $e=1.2$ ) (which is embedded in the in-house evaluation software) was used to compare various translations. The author herself conducted further qualitative analysis of the outputs. Hence, it was necessary to control the number of sentences included in the test. Many studies in the literature use 200 sentences in their test sets (Wang et al. 2007; Raybaud et al. 2009). Following this practice, a test set with 200 sentences (2889 words) was randomly selected from the preposition sample we compiled. The reference translation was extracted as well for later evaluation and comparison. We do not give a full report of the experiment set-up below because the approaches that proved promising will be expanded in the following chapters on a larger sampler where detailed set-up and evaluation results can be found.

#### **4.5.1 Testing of Domain-Specific UDs**

The purpose of the first test was to check the effectiveness of the domain-specific UDs. Examination of the English to Chinese user dictionary created by Symantec reveals that only a few entries related to prepositions are present. The main (English to Chinese) user dictionary contains 9,596 entries, grouped into six categories, i.e. general noun, proper noun, verb, adjective, adverb and sequence. The first four categories are common word categories in grammar. *Sequence* is defined by Systran as “those words and phrases (especially fixed expressions) that do not undergo linguistic analysis, but that are accepted ‘as-is’ for the final translation.”<sup>50</sup> An example from the user dictionary of Symantec is “DB release level”. Table 4.4 shows the specific number of each category and Figure 4.3

---

<sup>50</sup> Systran online support:  
<http://www.systran.co.uk/translation-support/important-information/dictionary-manager/dictionary-coding-user-guide#dcug121> [last visited 2010-05-15]

shows the proportions. In both cases, noun and proper noun are grouped under one category.

	Total entries	noun/proper noun	adjective	verb	sequence	adverb
Number	9596	8933	375	240	43	5

Table 4.4: Distribution of entries in the Symantec User Dictionary

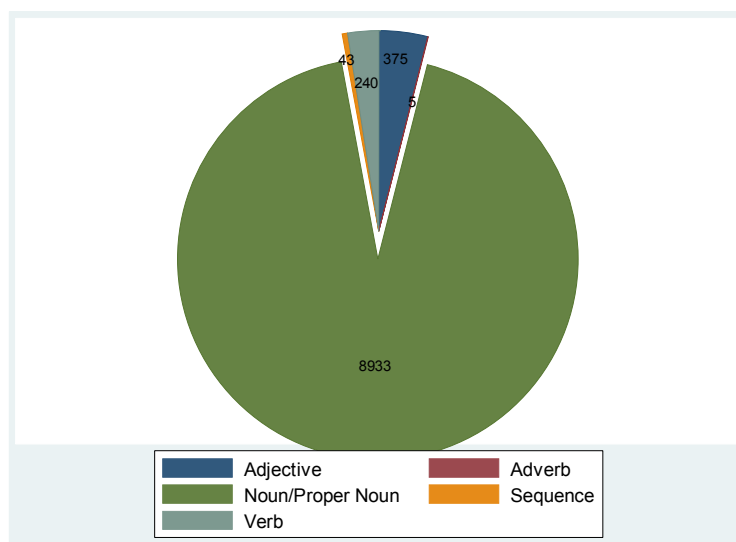


Figure 4.3: Distribution of entries in the Symantec User Dictionary

As can be seen, no separate category for prepositions is defined in the user dictionary. Within all entries, prepositions occur in less than 2% of all entries (about 180 entries containing prepositions, such as “job **in** progress”). The reason for having so few entries containing prepositions is due to the polysemous nature of prepositions and the fact that translations of prepositions are context-dependent as mentioned in Section 3.2.1.

The pilot test sample was first translated by Systran without the domain-specific UDs and then was translated again with the Symantec UDs. The GTM scores of the whole sample significantly improved from 30.62 (without domain-specific UDs) to 41.25 (with domain-specific UDs). However, while the overall translation becomes better (according to GTM), the number of preposition

errors remains the same in the two translations due to the small number of preposition entries. With the UD's of Symantec, the RBMT system can produce more accurate translation of specific term, but not necessarily more accurate translation of prepositions. Since this method exhibits potential, in Chapter 7, we report on how we extended this approach by proposing to extract a unique preposition dictionary automatically.

### **4.5.2 Testing of CL**

The second pilot test was on the effect of CL rules. We have stated that Symantec has a set of its own authoring rules for the purpose of MT. However, there is no rule regulating the use of prepositions. We gathered from the literature the following rules about the use of prepositions.

- (1) Rule 1: Verb + prep should be avoided and replaced with single word verbs. (cf. O'Brien 2003)
- (2) Rule 2: Avoid sentences ending with prepositions. (cf. O'Brien 2003)
- (3) Rule 3: Repeat prepositions in conjoined prepositional phrase where appropriate. (cf. Mitamura 1999)

Examining the Baseline translations revealed that sentences with four or more prepositions were usually translated incorrectly due to the complexity of the source sentence. In this exploratory test, we defined the fourth rule as follows:

- (4) Rule 4: Avoid using four or more prepositions in a sentence.

Note that this rule was based on the preliminary analysis of the test sample. The number of appropriate prepositions is difficult to specify and no standard or rule has been put forward.



We defined those rules in acrolinx IQ and then input our test sample to be checked. The flagged sentences were manually verified to ensure that no false alarms were included. We then rewrote those sentences and machine translated the sample once again.

For the 200 test sentences, only 8 sentences were flagged and were rephrased. The GTM scores of the two translations show no significant improvement, from 41.25 to 41.61. The number of errors in terms of translation of prepositions is also slightly reduced. The total number of errors can be found in Table 4.5 at the end of this section.

The biggest limitation of CL is that the rules have to be manually coded and sentences have to be manually rewritten. In addition, several testing phases are required in order to avoid degradations. Although only four rules on 200 sentences were tested, we opted not to further extend this method, due to the limited success of the pilot results.

### **4.5.3 Testing of Automatic S&R**

During the process of analysing the Baseline MT output, certain patterns appeared in those problematic translations of prepositions. One problem was incomplete circumposition translation. This could be partially corrected by global S&R through regular expressions (see Section 3.5 in Chapter 3). For example, if the words following *in* or *on* are proper nouns, translations of these prepositions were usually found to be incomplete in our sample, requiring a postposition to be added. The automatic S&R rule we proposed can be defined as follows:

If in the source sentence, there is a preposition *in* at the beginning of the sentence followed by a proper noun and a comma; while in the target sentence, the character 在 (*in*) is found followed by the translation of

the noun and then a comma, if character 中 is not present before the comma, add it.

However, the biggest problem again is that each rule has to be manually crafted. This is made worse by the fact that circumposition is preposition and context-dependent. For different prepositions, the postposition varies and a translation may be both correct with or without the postposition. Another problem is that while specific words, such as the English preposition *in* and Chinese word 在, could be defined easily, it is difficult, however, to describe the exact position of the postposition. Additionally, structural errors are also difficult to define and correct using this method.

As mentioned, Symantec has a post-processing module based on regular expressions which globally finds and replaces errors for many language pairs (Roturier et al. 2005); however, due to the complexity of Chinese prepositions as well as the limitations of S&R, the only rule specifically created for Chinese was to correct a punctuation problem. We did not find any general rules that could be applied from the pilot test. Hence, no evaluation results are reported for this approach.

It is worth noting that the potential of global S&R is also found in terms of source rewriting to create “pseudo” English that best suits the RBMT system (Aranberri 2009). However, the same limitations we just discussed also apply. We propose an alternative for automatically creating pseudo-English that better suits the RBMT system, i.e. a novel source pre-processing method through the use of an SMT system. This method will be introduced in Chapter 7.

#### 4.5.4 Testing of SPE

A statistical post-editing (SPE) system is built by feeding an SMT system a monolingual aligned parallel corpus: raw RBMT system outputs and corresponding human translated (or human post-edited MT output) reference translations, both of which are in the target language. The engine learns the changes between a raw MT output and a reference translation, calculates the probabilities of the changes, and will edit a new RBMT output based on the knowledge learnt. The importance of SPE has been reported widely (Roturier and Senellart 2008; Dugast et al. 2007; Simard et al. 2007a; 2007b).

Using our corpus compiled in Section 4.3 and following the instructions on the Moses website, we trained an SPE for the RBMT system using the Moses toolkit. We first translated the test set using Systran to get the raw RBMT system. Then, the SPE system was employed to post-edit this raw Systran output. The GTM score changed significantly from 41.25 (the Baseline translation) to 52.89 (the post-edited output). In terms of the number of errors, 25 errors were corrected.

In summary, Table 4.5 shows all the GTM scores of the approaches in the pilot tests and the number of preposition errors.

Method	GTM Score	No. of Preposition Errors
Systran	0.306	158
Systran+UDs (Baseline)	0.413	158
Systran+UDs+CL rules	0.416	153
Systran+UDs+SPE	0.529	133

Table 4.5: GTM scores and number of errors in the translations

The numbers in Table 4.5 make it clear that the translation of the test sample from SPE is the best according to the GTM scores and contains the least number of errors based on the analysis of the author. In order to further exploit the

benefits that an SPE system brings to the overall translation, this study goes a step further by controlling an SPE system to post-edit only the translation of prepositions. The detailed process of using the SPE system on a larger sample and the procedures for controlling the SPE system are all reported in Chapter 6.

## **4.6 Summary**

This chapter explains the methods we use to gain answers to the research questions. The core objective was to design the research following the major principles employed to address issues such as validity and reliability in order to obtain valid results.

The main MT systems employed were introduced first. This was followed by a justification of the forms of human evaluation and the suitable automatic evaluation metrics. For human evaluation, ranking outputs both at preposition-level and sentence-level is selected. Four professional translators were employed by Symantec to conduct the evaluations, complemented by three widely-used automatic metrics, namely, GTM, TER and BLEU. The corpus was prepared and a representative preposition sample was constructed.

Several pilot tests were carried out to assess the existing pre- and post-processing approaches, and how they might improve the translation of prepositions. UDs and statistical post-editing show greatest potential and were deemed to be worthy of further investigation while the other approaches were deemed to have less potential for this study and were therefore not pursued. However, before introducing the approaches, the first step is to report the overall translation quality of prepositions, i.e. are all prepositions problematic for the

RBMT system, etc. The next Chapter presents the first human evaluation with the aim of answering the first three research questions.

## **Chapter 5: Error Analysis of the Translation of Prepositions**

In Chapter 3, we outlined the main research question and five sub-questions. The first three sub-questions are: Which prepositions are translated incorrectly? Which of the five errors we outlined is the most frequent? And what is the most salient error associated with each preposition? In the pursuit of answers to these three questions, this chapter reports the results of a human evaluation on the translation of the representative preposition sample constructed in Chapter 4.

There are three sections in the current chapter. Section 5.1 introduces the experimental set-up. It begins by reviewing the preposition sample extracted and the error categories outlined. Next, the detailed preparation process of the evaluation is presented, including the evaluation sheets, the instructions and the questionnaires. Section 5.2 first examines the inter-evaluator correlation in order to check the reliability of the results. Based on the information gathered, the first three research questions are answered. Finally, Section 5.3 concludes this chapter.

### **5.1 Experiment Set-up**

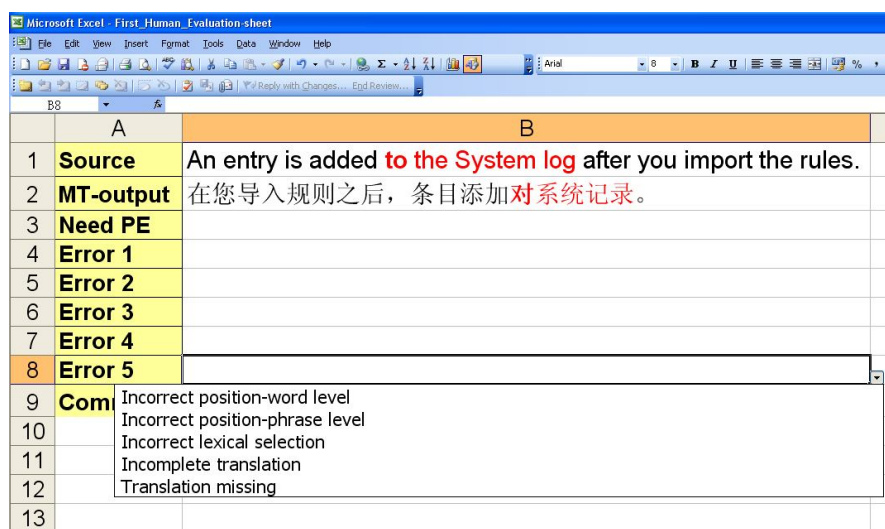
As mentioned in the chapter on methodology, a sample with 1000 prepositions was randomly selected and machine translated by Systran (the Baseline). Before applying any approach to improve this Baseline translation, it is necessary to review the overall quality of this Baseline translation, especially the translation of prepositions. To pinpoint the errors in translating prepositions, four professional translators were asked to examine the Baseline translation and select from the five basic error categories outlined in Chapter 3. These five errors are: Incorrect Lexical Selection (ILS), Incorrect Position-Word Level (IP-WL), Incorrect

Position-Phrase Level (IP-PL), Incomplete Translation (IT) and Translation Missing (TM). The human evaluators were also encouraged to note any new error types they encountered.

The instructions for the evaluation were written both in English and Chinese to ensure that the evaluators understood the requirements fully. There were two tasks in this evaluation: judging if a translation needed post-editing or not; and selecting the preposition-related errors present in a translation. To do the first task, evaluators were expected to be familiar with the overall quality of translation expected in the IT-domain. A second condition was that they should not be biased against MT technology and to believe that all MT output needs post-editing. To do the second task, evaluators were expected to have a sufficient grammatical knowledge of Chinese and English in order to select the appropriate error category. To gather the above-mentioned information, a questionnaire regarding their work experience, linguistic background, etc (see Appendix A) was administered. In addition to the questionnaire, instructions with error samples were also presented to the evaluators (see Appendix B). The 1000 sentences were ordered randomly in the evaluation sheet with one preposition/prepositional phrase highlighted in red. The corresponding Chinese translation of the prepositional phrase was also highlighted in the same colour. Although the evaluators were asked to focus on the highlighted sections, they were also required to read the whole sentence instead of evaluating the highlighted parts in isolation. This is especially important to pinpoint position errors of translations. Figures 5.1 and 5.2 show an example sentence in the final evaluation sheet.

1	Source	An entry is added <b>to the System log</b> after you import the rules.
2	MT-output	在您导入规则之后，条目添加 <b>对系统记录</b> 。
3	Need PE	
4	Error	Yes
5	Error	No
6	Error 3	
7	Error 4	
8	Error 5	
9	Comment	

Figure 5.1: Sample sentence in the evaluation sheet (error analysis-task 1)



	A	B
1	Source	An entry is added <b>to the System log</b> after you import the rules.
2	MT-output	在您导入规则之后，条目添加 <b>对系统记录</b> 。
3	Need PE	
4	Error 1	
5	Error 2	
6	Error 3	
7	Error 4	
8	Error 5	
9	Com	Incorrect position-word level
10		Incorrect position-phrase level
11		Incorrect lexical selection
12		Incomplete translation
13		Translation missing

Figure 5.2: Sample sentence in the evaluation sheet (error analysis-task 2)

Figure 5.1 shows the first evaluation task, i.e. ascertaining whether or not the translation of the highlighted preposition and/or prepositional phrase needs post-editing. The evaluators then selected the types of errors from the dropdown lists shown in Figure 5.2. The last column was used to record any comments the evaluators had, such as other types of errors. Two evaluators added some comments, both of which mainly suggested the correct translations for those incorrectly translated prepositional phrases. One evaluator pointed out in a few cases that redundant translations of prepositions were generated.



## 5.2 Evaluation Results

The following information was collated from the questionnaires. First of all, as stated, the four evaluators are professional translators who had worked on a wide range of projects in the IT-domain. The average number of words each translator has translated is around 3.5 million words. Therefore, they were considered sufficiently familiar with the general quality expectations for translation in the IT-domain. They all received an education in the grammar of both Chinese and English at high school and three of them have additional education at university. We conclude, therefore, that they are capable of judging the error types for translation. They all expressed willingness to work with MT systems, hence, their evaluation of MT output will not be biased by their opinion of MT. Based on this information, we can move on to seek answers for the questions we put forward at the beginning of this Chapter.

Section 2.2.3 of Chapter 2 pointed out that the inter-evaluator correlation (the Kappa score) should be calculated in order to indicate the reliability of human evaluation results. We used the Excel Kappa Calculator template (King 2004) to calculate Kappa scores through this study. This template is designed specifically for evaluation projects with multiple evaluators. In addition, it was an Excel template which greatly simplified the computing process for lay people. As all the data in this study were gathered in Excel and multiple evaluators were employed, this template is appropriate for this study. In addition, this template has been employed in similar research projects (e.g. Aranberri 2009). A snippet of how this calculator works is shown in Figure 5.3.

<b>Calculating a Generalised Kappa Statistic for Use With Multiple Raters</b>						
Calculations based on equations presented in Fleiss ( <i>Statistical Methods for Rates and Proportions</i> , 1981, pp. 229-232)						
<i>Directions: Enter values and data in shaded areas only.</i>						
Enter # of raters (m):	4					
Enter # of subjects (n):	1000					
# of categories (k):	2					
For each item below, enter the number of raters who placed the item into each respective category (delete/add rows as necessary):						
		<b>n of raters</b>				
	<b>Item</b>	<b>CAT1</b>	<b>x(m-x)</b>	<b>CAT2</b>	<b>x(m-x)</b>	<b>Sum x2</b>
	1	1	3	3	3	10
	2	2	4	2	4	9
	3	3	3	2	4	13
	⋮	⋮	⋮	⋮	⋮	⋮
	1000	3	3	1	3	10
*****						
gen kappa =	0.329					

Figure 5.3: Excel Kappa calculator template

The above form is a simplified form of the template with all the formulae hidden. The main information which needs to be input into the template includes: the number of raters (or the number of evaluators in this study), the number of subjects (or the number of sentences) and the number of categories (the number of options humans can choose). Take the first evaluation task of this study for example, i.e. whether the translation of prepositions requires post-editing or not, there are two categories (options) that humans can choose: Yes or No. Similarly, on the question of error selection, the options are also Yes (an error is present) or No (an error is not present). The specific Kappa scores related to each of the sub-questions listed above are reported in the following sections.

### 5.2.1 Question 1: Which prepositions are translated incorrectly?

The first question we need to answer is how many prepositions in the sample were translated unsatisfactorily according to the professional translators. The results can be separated into three groups. The first group contains prepositions whose translations were judged to be acceptable and not in need of further post-editing by at least three evaluators (Do not Need PE). In contrast, the second group

includes translations which the majority of the evaluators evaluated to be unacceptable (Need PE). And the third group includes prepositions for which no conclusive decision was made. Table 5.1 provides the number of prepositions in each of these groups.

Category	No. of Evaluator	No. of Preposition
Need PE	$\geq 3$	447
Do not need PE	$\geq 3$	448
Need PE vs. Do not Need PE	2 vs. 2 (ties)	105
	Total	1000

Table 5.1: Number of prepositions needing PE vs. those not needing PE

It can be seen from Table 5.1 that nearly half of the translated prepositions were judged as needing post-editing and another half not in need of post-editing. There are 105 prepositions out of 1000 where no majority vote was obtained. In other words, human evaluators agreed with each other most of the time (89.5%), indicating high inter-evaluator agreement. The inter-evaluator correlation ( $K$ ) for this evaluation task is  $K=.583$  (moderate agreement) which means that the four evaluators reliably agreed with each other on which translation needed post-editing.

For those prepositions in the group of “Need PE”, we further investigated the occurrence rate of these prepositions compared to their original numbers selected in the sample. Recall that only the top ten prepositions were included in this study and the number of each preposition selected in the sample was based on their frequency in the whole corpus (cf. Chapter 4). Table 5.2 reports the results.

Preposition	No. of Need PE	Total No. of occurrence	Percent needing PE
in	144	186	77.42%
on	88	121	72.73%
to	54	133	40.60%
for	52	160	32.50%
of	37	179	20.67%
from	31	70	44.29%
with	26	54	48.15%
by	14	39	35.90%
about	0	34	0%
as	1	24	4.17%
Total	447	1000	44.70%

Table 5.2: Percentages of prepositions being mistranslated

Table 5.2 shows that the more frequent a preposition is, the more often it was judged as needing post-editing. However, not all the top ten frequent prepositions are problematic for the RBMT system. For example, the translation of preposition *as* and, in particular, the translation of *about* was not seen to be problematic by the majority of the evaluators. On the other hand, over 70% of occurrences of *in* and *on* were translated unsatisfactorily by the RBMT system.

To sum up, the RBMT system can translate half of the prepositions correctly without further post-editing needed. One important question is “In what cases will a preposition be translated incorrectly?” One assumption might be that prepositions in shorter sentences tend to be less problematic than those in longer and complex sentences. However, an analysis of sentence length by the author showed that this assumption does not hold true. The average length of sentences in the “Need PE” group is 17 words per sentence while it is 16 words in the “Do not need PE” group. In addition, there are both simple and complex sentences in both groups. A difference found between the two groups of sentences is that there are more fixed prepositional phrases, e.g. *such as*, *for example* in the “Do not need

PE” group than in the other group. As mentioned earlier, fixed phrases are usually not ambiguous and the meaning is fixed and can be encoded in the dictionaries. Therefore, their translations are usually correct. On the other hand, not all the meanings of single prepositions can be encoded into dictionaries due to the polysemous nature of prepositions, as discussed in Chapter 3. This is the main explanation for the fact that around half of the prepositions are translated incorrectly. The fact that the translation of a preposition is entangled with the translation of other parts of a sentence and is dependent on context may be another reason.

### 5.2.2 Question 2: How frequent is each error?

The second question we want to explore is which error is the most frequent. For all the sentences evaluated, the numbers of each error category selected by the evaluators was counted. Table 5.3 presents the results.

Error	Occurrences
Incorrect Lexical Selection	758
Incorrect Position-Phrase level	552
Incorrect Position-Word level	78
Incomplete Translation (IT)	665
Translation Missing	157

Table 5.3: Number of each error assigned by the evaluators

We can see that overall the most frequent error selected is Incorrect Lexical Selection (ILS), followed by Incomplete Translation (IT) and then Incorrect Position at Phrase Level (IP-PL). The numbers in Table 5.3 were counted from the whole sample. Results in Table 5.1 show that translations of 447 prepositions still need post-editing according to the majority of evaluators. Since we are more interested in these prepositions, the distribution of errors among these prepositions was calculated. Figure 5.4 illustrates the total number of each error type selected

by the four evaluators on the 447 preposition instances. The most frequent error selected in the “Need PE” group is Incomplete Translation (IT), followed by Incorrect Lexical Selection (ILS).

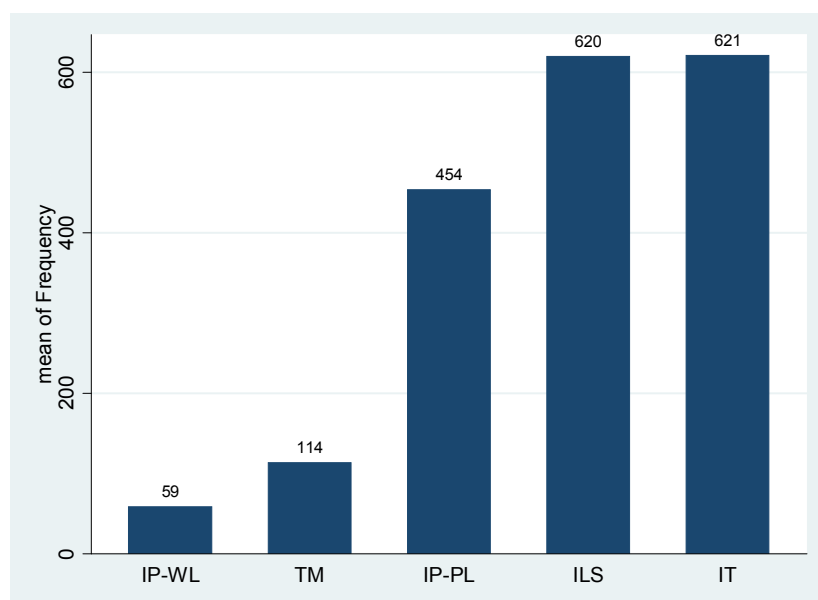


Figure 5.4: Error distribution among the 447 prepositions

We also checked the inter-evaluator agreement between the evaluators as to which error was present in each translation for the 447 preposition instances. The inter-evaluator agreement here refers to how consistent the evaluators were as to the type of errors in a translation. Table 5.4 shows the breakdown  $K$  scores between evaluators in respect of each error type.

Error	K value
Incorrect Lexical Selection	0.419
Incorrect Position-Phrase level	0.210
Incorrect Position-Word level	0.023
Incomplete Translation (IT)	0.554
Translation Missing	0.447
Average	0.476

Table 5.4: Inter-evaluator agreement for each error type

Overall there is a moderate agreement for the selection of error types, hence, in general, evaluators did not only agree with each other as to whether a translation

was acceptable or not, they also agreed with each other in general as to the type of errors present. This is confirmed by the fact that for 339 prepositions out of the “Need PE” group (447 prepositions in total), there is at least one error type that was agreed by at least three evaluators.

However, the agreement on Incorrect Position (IP), at both word level and phrase level, is extremely weak. This may suggest that the definition for Incorrect Position was not precise enough or that the division between phrase-level and word-level is not clear-cut. Another conjecture is that as evaluators have to look at the whole sentence to determine a position error which requires extra effort compared to lexical level errors, evaluators might tend to just focus on the lexical level and discard the position aspect. This may be the reason for the higher occurrence of Lexical Selection and Incomplete Translation over position errors.

### 5.2.3 Question 3: What are the errors pertaining to each preposition?

We also associated the errors with each preposition aiming at revealing which error occurs most frequently for a specific preposition (Table 5.5). Again, only the results for the 447 prepositions where at least three evaluators agreed on the quality are reported.

Preposition	ILS	IP-PL	IP-WL	TI	TM
as	<b>4</b>	0	0	0	0
by	22	<b>26</b>	3	0	1
with	<b>72</b>	32	4	2	3
from	31	<b>59</b>	2	5	0
of	24	35	11	6	<b>62</b>
for	<b>140</b>	48	5	2	21
to	<b>154</b>	41	10	4	14
on	74	89	12	<b>221</b>	4
in	99	114	10	<b>381</b>	10

Table 5.5: Distribution of errors in the translation of each preposition

Table 5.5 shows the distribution of errors of each preposition with the most frequent error in bold. It is apparent from Table 5.5 that the type of error is distributed differently for different prepositions. Recall that only one case of *as* was mistranslated and the error associated with *as* is Incorrect Lexical Selection (ILS). This error is also the most frequent error for prepositions *with*, *for* and *to*. Incomplete Translation (IT) is most often found in the translation of prepositions *in* and *on*. Position error is the biggest problem for preposition *from* and *by* while Translation Missing is prevalent in the translation of the preposition *of*.

### 5.3 Summary

From this evaluation, the following conclusions about the machine translation of prepositions from English to Chinese in this corpus can be drawn. First, while the RBMT system can produce satisfactory translations for almost half of all the prepositions, another half of the MT output for prepositions still need post-editing. Second, among the top ten frequent prepositions, some prepositions seem to be handled better (such as *about* and *as*). In addition, the most frequently occurring prepositions are not necessarily the most problematic ones. Third, the most common types of errors vary across prepositions. After quantifying the problems in translation of prepositions from English into Chinese and specifying the nature of the problem more closely, we can now move on to suggesting different methods for tackling these problems.

The major reason for the irregular occurrence of errors is due to the fact that English prepositions are polysemous and the translation correspondents are variable. Instead of trying to work on each preposition separately, this study proposes several approaches to work on prepositions in general. For example, to reduce the number of Incorrect Lexical Selections (ILS) and Incomplete



Translations (IT), a special preposition dictionary and a Statistical Post-Editing (SPE) approach are proposed. To correct the error of Incorrect Position, an approach to automatically re-writing the source into Chinese-flavoured English has been proposed. These approaches will be introduced in the following two chapters respectively.

## Chapter 6: Statistical Post-editing

In Chapter 4 we identified that Statistical Post-Editing (SPE) is one of the potential solutions for targeting some problems in translation of prepositions. Plitt and Masselot (2010) showed that SPE could greatly increase the productivity of translators in a typical localisation context. Results of the pilot test (see Chapter 4) concur with the statements of some previous studies that an SPE system can significantly improve the raw output of Systran (Roturier and Senellart 2008; Dugast et al. 2007).

On the other hand, SPE also generates degradations in the translation of prepositions (Dugast et al. 2007). Although it has been suggested that “adding a linguistic control mechanism” (Dugast et al. 2007: 223) could reduce the degradations of SPE, no controlled study, to the author’s knowledge, has been conducted. Hence, the first objective is to explore the effects of controlling the SPE module in such a way that it can focus more on post-editing the translation of prepositions. The principal aim of this test is to measure the potential improvements in the translation of prepositions if we control the process of SPE.

So far, SPE has been used as a general approach and no specific attention has been paid to what type of prepositions errors could be corrected by SPE. Therefore, our second aim in this chapter is to give an account of the errors that could be corrected by SPE based on the analysis of a larger sample.

This chapter is organised as follows. Section 6.1 begins by giving a brief introduction to the general experimental set-up. This is then followed by a step-by-step explanation of our proposal to modify the SPE module and the preparation of human and automatic evaluations. Section 6.2 examines the

reliability of human evaluation and the overall results are reported. Both qualitative and quantitative analyses of the translation of prepositions are conducted. In addition, sentence level evaluation was also conducted in order to measure the effect of the modified SPE on overall sentence level translation. Based on the findings of this test, the correlation between automatic and human evaluation at sentence level is scrutinised. Section 6.3 summarises this chapter.

## **6.1 Experiment Set-up**

As mentioned, the Baseline system for this study is Systran and its output is the Baseline translation to which other translation variants are compared. The SMT system used to build the SPE module is the Moses toolkit (Koehn et al. 2007). The basic mechanism of SPE is that an SMT system is trained using a set of raw RBMT system translations (the “source” sentences) and their corresponding reference translations (the “target” sentences), which are either a human post-edited version of the RBMT output or direct human translation of the original source text. These constitute a pair of monolingual parallel corpora. Using these corpora, the SMT system learns to “translate” (or “post-edit”) raw RBMT system output into better quality text.

Generally speaking, three corpora are needed: a training set, a tuning set and a test set. A training set contains monolingual parallel corpora that are needed to build the module. The purpose of a tuning set is to fine-tune the system in order to obtain the best possible performance. Finally, a test set is the sample text that is going to be translated and compared. Table 6.1 shows the sizes of the training, tuning and test set.

	# Sentences	# Words
Training set	5,951	84,349
Tuning set	944	15,884
Test set	944	15,916

Table 6.1: Preliminary training, tuning and test corpora

In SMT terms, the training corpus listed in Table 6.1 is a very limited corpus. This is due to the fact that only sentences containing one of the top ten prepositions were selected. Fortunately, in terms of training an SMT system for the purpose of SPE, it has been reported that even a small training corpus, could improve the performance of the RBMT system if the corpus is domain-specific (Roturier and Senellart 2008). Our pilot project (Section 4.5.4 in Chapter 4) has also shown that even with this small corpus, SPE could bring about significant improvement to the overall translation of the pilot sample (200 sentences).

The next section illustrates the process of building a basic SPE system and the process of modifying it in more detail.

### 6.1.1 Building and Modifying an SPE Module

There are four steps involved in building and modifying a basic SPE module. To make our explanation clear, a list of notations at each step was created. The notations for the corpora and translations are listed in Table 6.2.

	Meaning
$Train_{EN}$	The English training corpus
$Train_{ZH}$	The Chinese reference translation for the English training corpus
$Train_{MT}$	The Systran translation for the English training corpus
$Tune_{EN}$	The English tuning corpus
$Tune_{ZH}$	The Chinese reference translation for the English tuning corpus
$Tune_{MT}$	The Systran translation for the English tuning corpus
$Test_{EN}$	The English test sample
$Test_{ZH}$	The Chinese reference translation for the English test sample

Table 6.2: Notations for corpora used for SPE

- Step 1 Train and tune an SPE system

The SPE system required  $Train_{MT}$  (the “source” language) and  $Train_{ZH}$  (the “target” language) for training and  $Tune_{MT}$  and  $Tune_{ZH}$  for fine-tuning. The phrase table in the obtained SPE system is monolingual (Chinese) and contains raw RBMT output on one side and the reference translation on the other. Phrase tables are of vital importance to an SPE module (and to an SMT system) (cf. Chapter 2). An SPE module attempts to select the translations with the highest probability using its phrase table (which determines the accuracy of translations) together with a pre-extracted target language model (which determines the fluency of translations). Thus, the more precise and correct the phrase table, the higher the quality of the SPE output (cf. Chapter 2). The notation for the phrase table of this pre-trained SPE system is presented in Table 6.3.

Notation	Meaning
$Phrase_{MT \rightarrow REF}$	Monolingual phrase table containing phrases learnt from the raw RBMT output and the reference translations.

Table 6.3: Notation for the monolingual phrase table of the SPE module

- Step 2 Translate the test sample with the SPE module

To use the pre-trained SPE module, the test sample was first translated by Systran into Chinese. This is the Baseline translation to which other translation versions are compared. Next, the pre-trained SPE module was initiated to post-edit the raw Baseline translation to get a second version of the translation. This translation variant is called the default output of the SPE module as no modification was applied to the SPE system. The notations are shown in table 6.4 below.

Notation	Meaning
Baseline	The Systran output of the English test sample
SPED	The default translation of the SPE module which is obtained by post-editing the Baseline translation using the basic SPE system

Table 6.4: Notations for Baseline and SPED

The first two steps are depicted in Figure 6.1.

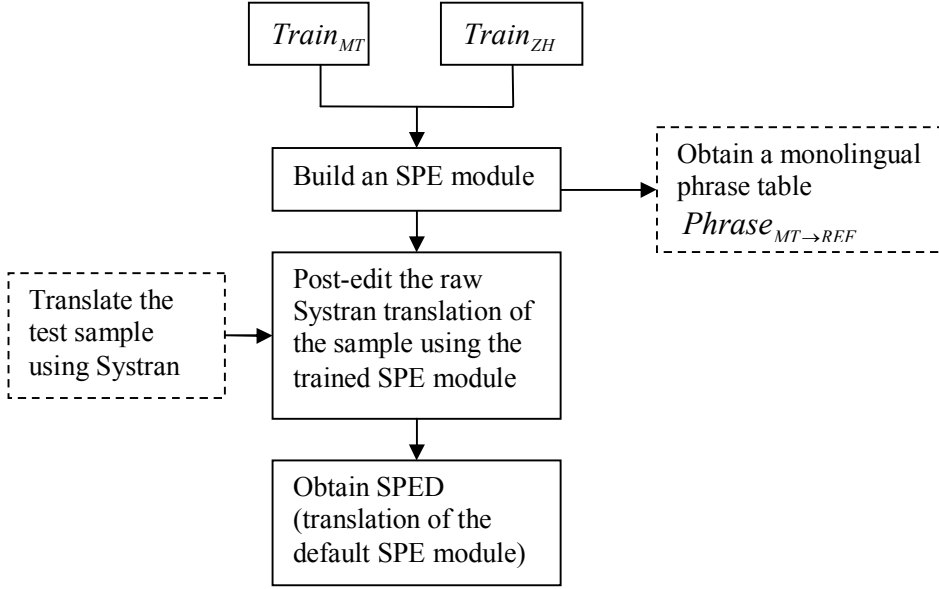


Figure 6.1: Flowchart of the first two steps in the process of modifying SPE

- Step 3 Modify the SPE system

This step involves modifying the core component (i.e. the phrase table) of this SPE module by removing phrases not containing prepositions. However, as mentioned, the phrase table ( $Phrase_{MT \rightarrow REF}$ ) in the unmodified SPE module is monolingual, with raw RBMT Chinese output on one side and reference Chinese translation on the other side. It is necessary to find out which of the raw RBMT output strings were translated from English phrases with prepositions. In other words, we need the translation phrases between the source English and the raw RBMT output.

- Step 4 Generate a bilingual phrase table

Using  $Train_{EN}$  as the source language and  $Train_{MT}$  as the target language together with the statistical phrase toolkits of Moses, we obtained a bilingual phrase table. The resulting phrase table contains pairs with English phrases on one side and raw RBMT output on the other. Next, any phrase pairs where the English side contained no prepositions were removed. The resulting phrase table (which is a preposition phrase table to be more specific) will help us to modify the default SPE system obtained in the following steps. The notations used at this step are shown in Table 6.5. The phrase table from  $Phrase_{EN \rightarrow MT}^{prep}$  was used to remove phrases that do not relate to prepositions in  $Phrase_{MT \rightarrow REF}$  obtained in step 1.

Notation	Meaning
$Phrase_{EN \rightarrow MT}$	Bilingual phrase table containing all possible corresponding translation sequences learnt from the English training data and the RBMT translation
$Phrase_{EN \rightarrow MT}^{prep}$	Phrase table with English phrases containing prepositions and their corresponding RBMT translation

Table 6.5: Notations for bilingual and preposition phrase table

Comparing  $Phrase_{EN \rightarrow MT}^{prep}$  (from step 4) and  $Phrase_{MT \rightarrow REF}$  (from step 1), we can see that the common part between these two phrase tables is the raw RBMT output.  $Phrase_{EN \rightarrow MT}^{prep}$  contains English phrases with prepositions and their corresponding **raw RBMT Chinese translations**.  $Phrase_{MT \rightarrow REF}$  contains **raw RBMT Chinese translations** and the corresponding reference Chinese translations. For example, In  $Phrase_{EN \rightarrow MT}^{prep}$ , the following phrases are present:

English phrase	Raw RBMT translation
In the Requirement tab	要求 表 里 [gloss: Requirement tab in]

And in  $Phrase_{MT \rightarrow REF}$ , the following phrases are found:

Raw RBMT translation	Reference translation
要求 表 里	“ 要求 ” 表 中
[gloss: Requirement tab in]	[gloss: “Requirement” tab in]

Therefore, the two phrase tables can be connected through the raw RBMT translation as follows:

$Phrase_{EN \rightarrow MT}^{prep}$		$Phrase_{MT \rightarrow REF}$	
English phrase		Raw RBMT translation	Reference translation
In the Requirement tab		要求 表 里	“ 要求 ” 表 中

We compared  $Phrase_{MT \rightarrow REF}$  to  $Phrase_{EN \rightarrow MT}^{prep}$  and retained those phrase pairs in  $Phrase_{MT \rightarrow REF}$  where the raw RBMT side in  $Phrase_{EN \rightarrow MT}^{prep}$  could be matched to  $Phrase_{MT \rightarrow REF}$ . However, we are aware that not all the raw RBMT phrases in  $Phrase_{EN \rightarrow MT}^{prep}$  can be found in  $Phrase_{MT \rightarrow REF}$ . Even if a phrase is found in both phrase tables, there are two types of matches. Let us continue with the same example to illustrate this point. Suppose the following phrase is present in the  $Phrase_{EN \rightarrow MT}^{prep}$ :

English phrase	Raw RBMT translation
In the Requirement tab	要求 表 里 [gloss: Requirement tab in]

And in  $Phrase_{MT \rightarrow REF}$ , we may find two matching phrases:

1) Raw RBMT translation	Reference translation
要求 表 里	“ 要求 ” 表 中
[gloss: Requirement tab in]	[gloss: “Requirement” tab in]



2) Raw RBMT translation	Reference translation
将 名称 填 在 要求 表 里 [gloss: name fill in Requirement tab in]	在 “ 要求 ” 表 中 填入 名字 [gloss: In “Requirement” tab in fill name]

The first match indicates that the whole phrase from  $Phrase_{EN \rightarrow MT}^{prep}$  can be fully matched in  $Phrase_{MT \rightarrow REF}$ . The second match indicates that the phrase from  $Phrase_{EN \rightarrow MT}^{prep}$  may be contained as a part of a phrase in  $Phrase_{MT \rightarrow REF}$ . We called the first match a Full Match, i.e. a phrase in  $Phrase_{EN \rightarrow MT}^{prep}$  is equally and exactly matched in  $Phrase_{MT \rightarrow REF}$  and the second match a Partial Match, i.e. a phrase in  $Phrase_{EN \rightarrow MT}^{prep}$  is matched into part of a phrase in  $Phrase_{MT \rightarrow REF}$ . The difference between the two matches is that the latter (case 2) contains extra information that is not necessarily related to prepositions. Using the first match can minimise this unrelated information which may cause degradation in the translation of prepositions. The problem is that phrases like the one in case 2 would be missed although it did contain translation of prepositions. Using the second match can ensure that all phrases related to prepositions are included but faces the challenge of including contexts not related to prepositions. Based on these two matches, we filtered the  $Phrase_{MT \rightarrow REF}$  in two ways. The first way is to only keep phrase pairs that have a full and exact match between  $Phrase_{MT \rightarrow REF}$  and  $Phrase_{EN \rightarrow MT}^{prep}$  (match 1). This removed 76.9% of the phrase pairs from  $Phrase_{MT \rightarrow REF}$ . The second way is to keep phrase pairs in  $Phrase_{MT \rightarrow REF}$  if it contains or is exactly matched to a phrase in  $Phrase_{EN \rightarrow MT}^{prep}$  (both match 1 and

match 2). In contrast to the first filtering, just 2.6% of phrase pairs in  $Phrase_{MT \rightarrow REF}$  were removed.

After filtering the phrase table of the general SPE system, two new SPE systems were generated. The Baseline translation was then post-edited again by each of the two new SPE systems and two new translations were obtained (Table 6.6).

Notation	Meaning
SPEP	Translation from the modified SPE module with the phrase table that was filtered based on Partial Matches
SPEF	Translation from the modified SPE module with the phrase table that was filtered based on Full Matches

Table 6.6: Notations for SPEP and SPEF

Figure 6.2 illustrates the above steps.

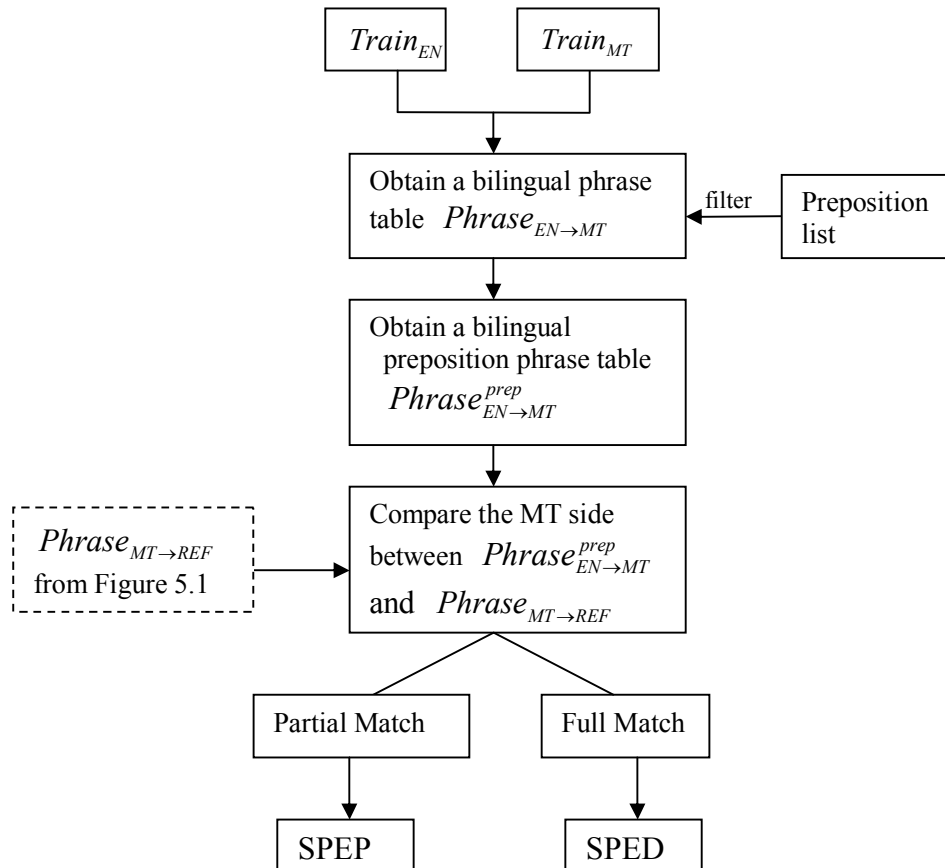


Figure 6.2: Flowchart of steps 3 to 5 in the process of modifying SPE

To reiterate the purpose of this test, we wanted to compare the Baseline, SPED, SPEP and SPEF, while focusing on analysing the gains and losses of modified SPE modules in particular for the translation of prepositions. To ascertain the level of gains or losses, a comparison and evaluation of these translations was conducted manually and automatically.

### 6.1.2 Evaluation Preparation

With regard to human evaluation, two evaluation tasks were designed in this project. The first evaluation is the preposition-level evaluation focusing on translation of prepositions. One English prepositional phrase and all the corresponding translations in the translation variants were highlighted in red. As in Chapter 5, four professional translators who are native speakers of Chinese were asked to rank the outputs from best (1) to worst (4). The translation variants were mixed and randomly arranged without showing which system they were from. Figure 6.3 is a snippet of the evaluation sheet on translation of prepositions. In addition to the evaluation sheets, instructions written both in English and Chinese (Appendix C) and a questionnaire (Appendix D) were also attached.

ID	Sentences	Ranking	Comment
Source	How protection is updated <b>on unmanaged clients</b>		
Reference	<b>非受管客户端</b> 更新 防护 功能 的方式		
* Output 1	保护怎么 <b>在未处理的客户端</b> 更新	<b>select here</b>	
* Output 2	保护 方式 <b>在非受管客户端</b> 更新	<b>select here</b>	
* Output 3	防护 再进 行 <b>非受管客户端</b> 更新	<b>select here</b>	
* Output 4	防护 怎么 <b>在非受管客户端</b> 更新	<b>select here</b>	

Figure 6.3: Screenshot of preposition level evaluation

The basic form of the evaluation follows the practice of the large-scale WMT evaluation (Callison-Burch et al. 2009; 2008; 2007). For each English sentence (“Source” in Figure 6.3), one reference translation (“Reference”) and its machine translations (“Output1”, “Output2”, “Output3”, and “Output4”) were provided.

Human evaluators were reminded that the reference was provided to help them understand the structure of the source sentence quickly and should not be considered as the only correct translation.

In addition to evaluating preposition translation, a sentence level evaluation was also conducted. The purpose of this evaluation is to find out if the modified SPE led to any degradation on the sentence level. The method of human evaluation at sentence level is the same as the evaluation at preposition level except that no constituents were highlighted (see Figure 6.4) and evaluators were asked to consider the entire sentence. The same evaluators were used for the sentence-level evaluation. Again, they were asked to rank all the outputs of a sentence based on their understanding of the source sentences.

ID	Sentences	Ranking	Comment
Source	For example, a rule may state that remote port 80 is allowed to the IP address 192.58.74.0, between 9 AM and 5 PM daily.		
Reference	例如，某规则可能说明，每天上午9点至下午5点之间，允许IP地址192.58.74.0访问远程端口80。		
* Output 1	例如，规则也许阐明，远程端口80每日允许对在上午9点和下午5点之间的IP地址192.58.74.0。	select here	
* Output 2	例如，规则可能指出远程端口80每天允许至上午9点下午5点之间的IP地址192.58.74.0。	1 2 3 4	
* Output 3	例如，规则可能指出远程端口80每天上午9点，并允许至点之间的IP地址192.58.74.0。	select here	
* Output 4	例如，规则可能阐明，远程端口80每天允许对在上午9点和下午5点之间的IP地址192.58.74.0。	select here	

Figure 6.4: Screenshot of sentence level evaluation

The test sample contains 1000 instances of prepositions in 944 unique sentences. To evaluate all the 1000 unique preposition instances at preposition level and then again all the 944 sentences, each of which has four outputs to compare (i.e. each evaluator has to read nearly 2000 English sentences and nearly 8000 MT outputs) was beyond the scope of time and budget available for this part of the project and so the number of sentences to be evaluated had to be reduced.

First of all, translations from all four systems that were identical were removed. This removed 31 sentences from sentence level evaluation (944 in total) and 475 sentences from preposition evaluation (1000 in total). In other words, while at sentence level, the translations of the four systems are different in most cases, the translation of prepositions is the same in 47.5% of the cases.

Furthermore, we also removed sentences from the evaluation if the outputs from SPED, SPEF and SPEP were the same. The major aim of the experiment was to establish whether modified SPE modules can produce better translation of prepositions than the unmodified SPE module compared to the Baseline translation, which could not be ascertained if the three translations were the same. It was observed that the three SPE modules share a lot of translations especially translations of prepositions in common. After filtering the test sample, 125 sentences remained to be evaluated at preposition level and 570 sentences at the sentence level. Note that after the deletion of duplicate translations, the four outputs were not always presented for evaluation. For example, for one sentence, only two translations were presented (e.g. Baseline and SPED) because SPEF and Baseline had the same translation and SPED and SPEP had the same translation.

The drawback of removing those duplicate translations was that data regarding intra-evaluator agreement (the consistency of an evaluator judging the same translation more than once) would be lost. However, as the central issue is to check which translation is better, a satisfactory level of inter-evaluator agreement (the consistency between all the evaluators) was considered to be enough to ensure that the final ranking of the outputs was valid.

Three automatic evaluation metrics were selected in the chapter on methodology as a complementary way of measuring the difference between

various translations. As a reminder, the scores of GTM and BLEU range from 0 to 1 and there is no maximum limit for TER scores. The higher a GTM/BLEU score, the better an MT output is; however, the lower a TER score, the better the MT output.

## **6.2 Evaluation Results**

Although the number of sentences to be evaluated has been reduced by removing duplicate translations, it still took each of four translators 24 hours to complete the evaluation task. All translations (125 sentences) were compared and ranked by the four translators at preposition level. However, four sentences (out of 570) were left unranked by one evaluator at sentence level evaluation. As no majority vote for these four sentences was achieved by the remaining three evaluators, these four sentences were deleted from the final results leaving 566 sentences to be examined at sentence level. H1, H2, H3 and H4 in the following sections represent the four human evaluators. The four translations are designated Baseline, SPD, SPEP and SPEF respectively.

### **6.2.1 Reliability of Human Evaluation**

To infer conclusions from the human evaluation, it is important to check the validity of the evaluation results in the first instance. The agreement level between the human evaluators and the Kappa correlation of the four evaluators were calculated. The common practice in analysing ranking results is to first expand the results into pair wise comparison (Callison-Burch et al. 2009; 2008; 2007). There are 2020 pairs of translations evaluated at sentence level and 217 pairs at preposition level. The overall agreement level was then obtained by counting the number of pairs where at least three evaluators agreed on the quality of the

translations divided by the total number of pairs. The Kappa multiple raters' inter-evaluator correlation was calculated using the Excel Kappa Template explained in Chapter 5. There are three options for this ranking task, i.e. for any pairs of comparison, the proportion of time that the evaluators judged one system as better than, worse than or equal to the other system. A break-down pair wise agreement level and Kappa values are presented in Table 6.7 for sentence level evaluation and Table 6.8 for preposition level evaluation.

	AGREEMENT LEVEL	KAPPA
H1-H2	68%	0.397
H1-H3	23%	-0.177
H1-H4	44%	0.147
H2-H3	18%	-0.256
H2-H4	42%	0.108
H3-H4	64%	0.293
Overall	44%	0.273

Table 6.7: Agreement level and inter-evaluator correlation (sentence)

	AGREEMENT LEVEL	KAPPA
H1-H2	70%	0.444
H1-H3	39%	0.081
H1-H4	62%	0.382
H2-H3	37%	0.044
H2-H4	61%	0.346
H3-H4	50%	0.237
Overall	63%	0.276

Table 6.8: Agreement level and inter-evaluator correlation (preposition)

Overall, the agreement level and correlation score is marginally higher for preposition evaluation than for sentence evaluation. At preposition level, there is a fair correlation ( $K=0.276$ ) between the evaluators and it is the same at sentence level evaluation ( $K=0.273$ , fair correlation). The agreement levels show that for sentence evaluation, the majority of evaluators reached agreement on less than half of the translations. In terms of judging translation of prepositions, the majority of evaluators reached agreement on 63% of the translations.

Given the large number of pairs being evaluated, both the overall K scores are significant at  $p < 0.01$ . That being said, a mere *fair* agreement is undoubtedly a low agreement level. However, this is a recognised problem in the MT research community and the correlation value constantly reported in the literature is also fair agreement (see Callison-Burch, et al. 2009; 2008; 2007).

Methods have been proposed to increase the inter-evaluator agreement such as discarding the judgement of evaluators who had the lowest agreement with others (Callison-Burch, et al. 2009). From Tables 6.7 and 6.8 we can see that H3 has the lowest correlation with the others. H1 and H2 do not have positive correlation with H3 even at sentence level. Removing the results of H3 improves the inter-evaluator correlation noticeably from 0.276 to 0.392 at preposition level evaluation and from 0.273 to 0.381 at sentence level. Nonetheless, the correlation still belongs to fair agreement. Discarding the results of human evaluation is not always the best solution, especially in cases where only a limited number of evaluators were employed. In addition, it would be more convincing to attempt to improve the inter-evaluator correlation before the evaluation rather than discarding the results after the evaluation which implies a waste of both time and budget. An important contribution of the study is that in Section 6.2.4 we propose a new approach to increase the inter-evaluator agreement in order to obtain more reliable evaluation results in our later evaluation experiments.

For the results presented below, all judgements have been retained. The next section moves on to examine the core research question of this evaluation, that is, whether the modified SPE module can produce better translation of prepositions.



## 6.2.2 Results - Translation of Prepositions

The first set of analyses compared the translations of prepositions between the SPE modules and the Baseline system. Two questions were addressed: first, which output was most preferred by the evaluators? And second, what were the errors that could be corrected by SPE?

### 6.2.2.1 Quantitative Analysis

The simplest and quickest way to check the difference between the four translations is to count the number of No.1 ranks that each translation was assigned by the evaluators. Recall that due to the deletion of shared translations, the four outputs were not always present in a sentence. Therefore, we expanded the number of outputs in all sentences into four, i.e. Baseline, SPED, SPEF, SPEP so that the number of translations was the same. The rankings from evaluators were repeated for those duplicated translation. Table 6.9 reports the results.

Output ID	Top-ranked frequency	Proportion (out of 125)
Baseline	37	29.6%
SPEF	50	40.0%
SPEP	63	50.4%
SPED	68	54.4%

Table 6.9: Number and percent of times that each system was top-ranked (preposition level)

Table 6.9 shows the number of sentences (and the proportion of times out of the total 125 sentences) that a system was ranked as #1 by at least three evaluators in preposition evaluation. The results reveal the most preferred output for human evaluators. We can see from the numbers that the translations generated by the unmodified SPE (SPED) is the most frequently preferred translation, with the Baseline translation being least preferred. The difference between the Baseline

translation and the SPED translation is noticeable but the difference between SPED and SPEP is relatively small.

An alternative method of comparison for the four systems is to check the average number of times that each system was judged as better than any other system (Callison-Burch et al. 2007; 2008; 2009). The results can reflect among all the translations evaluated, which one was evaluated as better for most of the time.

Table 6.10 shows the results.

Baseline	SPEF	SPEP	SPED
22.90%	28.90%	31.10%	32.10%

Table 6.10: The percent of times that a system was judged to be better than any other system (preposition level)

The results confirm the finding displayed in Table 6.9, i.e. the translation of SPED is better than any other system, most of the time. The translation generated by SPEP is slightly less preferred compared with SPED and the Baseline system is least preferred.

To test if the differences between different translations were generated by chance or can be considered as statistically significant, we applied the non-parametric Wilcoxon signed-rank test (Boslaugh and Watters, 2008; Woods et al, 1986). For any two systems, their rankings were extracted and compared through a statistical tool (SPSS). In Table 6.11, “>” indicates the column system is significantly better than the row system. “<” indicates the column system is significantly worse than the row system. And “≈” indicates there is no significant difference between the two systems. \* indicates statistical significance at  $p<0.05$  and \*\* at  $p<0.01$ . The lower the  $p$  value, the more significant the difference.

	Baseline	SPED	SPEP	SPEF
Baseline	/	>**	>**	≈
SPED	<**		≈	<*
SPEP	<**	≈		<*
SPEF	≈	>*	>*	

Table 6.11: Pair wise significance test (preposition level)

From the statistics in Table 6.11 we can conclude that SPED and SPEP are significantly better than the Baseline (at  $p < 0.01$ ) and than SPEF (at  $p < 0.05$ ). There is no significant difference between SPED and SPEP. The difference between the Baseline and SPEF is not significant either.

The reason that the translation quality of prepositions between SPED and SPEP (and between the Baseline and SPEF) is not significantly different from each other is mainly due to the fact that most of the translations from the two systems are the same. To demonstrate this, the number of identical translations for the highlighted parts of prepositions and prepositional phrases shared by each pair is listed in Table 6.12 with the percentages (out of 125 sentences) shown in parenthesis.

	SPED	SPEP	SPEF
Baseline	13 (10.4%)	17 (13.6%)	44 (35.2%)
	SPED	89 (71.2%)	16 (12.8%)
		SPEP	15 (12%)

Table 6.12: Number and percentage of shared translations by any two systems (preposition level)

Table 6.12 shows that SPED and SPEP share 71.2% of translations of prepositions in common (more than half of the translations of prepositions evaluated are the same between SPED and SPEP). Baseline and SPEF also share some translations in common although the number is much smaller compared to that of SPED and SPEP.

The key reason for the shared translations is the filtering process for the phrase table in Section 6.1.1. As mentioned, the way of obtaining the phrase table for SPEP (Partial Match) resulted in a phrase table that has 97.4% of the original phrases in SPED. Hence, the performance of the two SPE modules was almost the same. On the other hand, due to the small phrase table of SPEF (which is 23.1% the size of the phrase table of SPED obtained through Full Match), many raw Baseline translations were not post-edited. That is why the output from SPEF is not greatly different from the Baseline translation. Note that only translations of prepositions were taken into consideration; the situation is different when the whole sentence was examined (see Section 6.2.3 below).

In summary, in terms of translation of prepositions, the unmodified SPE system is significantly better than the Baseline translations and translations of the modified SPE systems. However, there are some Baseline translations which are better than translations from any SPE system (see Table 6.9), which means that in some cases after statistical post-editing, the translation quality suffered.

#### **6.2.2.2 Qualitative Analysis**

The original purpose of constraining and modifying the general SPE system was to increase the number of correct translations of prepositions. To check this, we compared each output from any SPE module with the Baseline translation. If the translation of the SPE was judged as better than the Baseline by at least three evaluators, it was defined as an improvement. If the translation of the SPE was judged as worse than the Baseline by at least three evaluators, it was defined as degradation. The translations were equivalents if they were judged as equal with Baseline translation by the majority of the evaluators. As can be seen from Table

6.13, SPED still has the biggest number of improvements, compared to the constrained SPEP and SPEF.

	Improvements	Degradations	Equivalents	Improvement/Degradation Ratio
SPED	55	22	32	2.50
SPEP	52	23	30	2.26
SPEF	26	18	31	1.44

Table 6.13: Improvement and degradation of each system compared to the Baseline

As stated above, translations of SPEP and SPED share many translations with each other. Therefore, the improvements brought about by SPEP and SPED are almost the same. Further examination shows that although there are 52 improvements brought about by SPEP, only 9 improvements do not overlap with the improvements of SPED. This again confirms that there is no significant difference between SPED and SPEP. On the contrary, although SPEF is slightly better than the Baseline translation, 77% of all the improvements are unique improvements that were not found in SPED. Hence, although SPEF failed to bring about the same number of improvements as SPED, it did bring about many unique improvements.

The results show that the constrained SPEP is not a successful modification compared to the unmodified SPE module (SPED) as there is no significant difference between them and fewer improvements were generated. Likewise, SPEF is also not a successful attempt. However, SPEF is worthy of more research as it produces many unique improvements. If the improvements of SPED and SPEF can be combined together, then the overall improvement ratio could be increased substantially. Overall, the general SPED which had the most phrases (preposition related or not) produced the best translation of prepositions.

Detailed linguistic analysis of general SPE output can be found in previous studies. For example, Dugast et al. (2007) reported the linguistic improvements and degradations of SPE on French translation. Roturier and Senellart (2008) analysed French, Japanese, German and Chinese outputs. Tatsumi and Sun (2008) compared the effect of SPE on Japanese and Chinese translations. Most studies include many linguistic categories in their analyses. For example, Roturier and Senellart (2008) reported the effect of their SPE experiment on 15 linguistic categories.

However, the research mentioned above has not examined the translation of prepositions in detail. In this study, we narrow down our qualitative analysis on the translation of prepositions from general SPED and compare it to the Baseline translation in order to reveal which errors in the translation of prepositions can be corrected by SPE. Since the general SPE has been employed in production by Symantec, the findings would be of practical use to the users/researchers in this context.

Translations generated by SPED and the Baseline systems were extracted separately, along with the human evaluation results. Only those sentences where at least three evaluators agreed on the quality of the translation were analysed. Sentences where the translation of SPED was evaluated as better than the Baseline were extracted. Using the error typology we set up in Chapter 3, the author conducted an analysis of the number of errors corrected by the SPED system.

It was observed that the most frequently corrected error is Incorrect Position (both word level and phrase level) which accounts for 47.37% of all the corrections. This is followed by Incomplete Translations (18.8%). The remaining

corrections are on lexical selection errors and errors related to missing translation. Examples of the most frequent corrections are presented below.

SPE can correct some position errors generated in the Baseline translation of prepositions, especially *of* phrases, see example (6.1). The gloss of the prepositional phrase is in brackets at the end of the translation.

---

Example 6.1

**Source:** Beside the **type of log** that you want to view, click View Logs and then click the name of the log.

**Ref:** 在您要查看的**日志类型**旁边，单击“查看日志”，再单击此日志的名称。[Gloss: log type beside]

**Baseline:** 在您想要查看**日志**旁边的请，**种类**单击视图日志然后单击这本日志的名称。[Gloss: log beside type]

**SPED:** 在您要查看的**日志**的**类型**旁，单击“查看日志”，然后单击“此日志的名称”。[Gloss: log's type beside]

---

The coloured words and glosses show that SPED improves the Baseline translation (Systran output) by correcting the position error of the prepositional phrase *the type of log*. The correct order should put the translation of *log* directly before the translation of *type*.

An example of SPE correcting incomplete translation errors especially for prepositions *in* and *on* (the translation of which usually requires circumpositions in Chinese) is shown in Example 6.2 on below.

---

### Example 6.2

**Source:** On the Windows XP taskbar, click Start Control Panel.

**Ref:** 在 Windows XP 任务栏 上 ， 单击 “ 开始 ” “ 控制面板 ” 。

**Baseline:** 在 Windows XP 任务栏 ， 请 单击 启动 控制 面板 。

**SPED:** 在 “ Windows XP 任务栏 上 ， 单击 “ 开始 ” “ 控制面板 ” 。

---

Comparing the translation of the highlighted prepositional phrase *On the Windows XP taskbar*, we can see that SPED adds the post-preposition 上 at the end of the phrase, thereby completing the meaning of the phrase and making the translation more fluent.

SPE is not without problem, however. Some of the translations after SPE become worse than the Baseline translation. For example, in example 6.3, the translation of preposition *on* was missing after SPE. However, overall there are more improvements than degradations brought about by the general SPED to the Baseline translation.

---

### Example 6.3

**Source:** Enter data on each panel to create the type of rule you selected.

**Ref:** 在 每个 面板 输入 数据 ， 以 创建 所选 规则 的 类型。 [Gloss: on each panel enter data]

**Baseline:** 输入 在 每个 面板 的 数据 创建 您 选择 规则 的 种类 。  
[Gloss: enter on each panel's data ]

**SPED:** 键入 每个 ” 面板 的 数据 创建 您 选择 规则 的 类型 。  
[Gloss: enter each panel's data]

---

To sum up, as in the pilot test in Chapter 4, the experiment on a larger sample conducted in this chapter concurred with the finding that SPE can greatly improve



translation of prepositions. Detailed linguistic examination shows that the general SPE is capable of correcting some position errors and incomplete translation errors. The proposal to modify an SPE system failed to bring more benefits to the translation of prepositions and translation of sentences.

Recall that there are two tasks in this evaluation, preposition evaluation and sentence evaluation. Having examined the effects of modified SPEP/F at preposition level, we now move on to review the translation of sentences.

### 6.2.3 Results - Translation of Sentences

The above analysis has shown that the modified SPE modules did not result in better translation of prepositions. This section examines the effect of the modified SPE systems at a sentence level.

#### 6.2.3.1 Automatic Evaluation Results

Three automatic metrics were applied to assess the difference between all translations. These metrics are GTM, TER and BLEU. Table 6.14 shows the automatic evaluation scores of each translation.

	GTM (e=1)	GTM (e=1.2)	BLEU	TER
Baseline	0.415	0.346	0.232	0.547
SPEF	0.520	0.437	0.374	0.436
SPEP	0.547	0.463	0.406	0.408
SPED	0.552	0.467	0.412	0.402

Table 6.14: Automatic evaluation scores of Baseline and SPEF/P/D

According to the scores, overall SPED is the best translation and the Baseline is the worst. There is just a slight difference between SPED and SPEP. To conduct significance tests on difference is complicated as we only have one test sample. For example, it is not possible to test if the differences between BLEU scores for Baseline and SPED are significant using only the two scores. Koehn (2004)

proposed a bootstrapping re-sampling method to solve this problem. Basically, to verify if the difference between two BLEU scores of two documents (say  $BLEU_A$  for text A and  $BLEU_B$  for text B) is significant or not, the same texts (text A and text B) can be randomly re-sampled a number of times to get a sufficient number of texts (such as  $A_1, A_2$ , etc. and  $B_1, B_2$ , etc.) for a statistical test. The new texts are then scored to get the BLEU scores and it is then ascertained whether the difference between these two sets of scores is significant or not. Using this method, we found that according to all automatic metrics, SPED is significantly better than SPEP at  $p < 0.05$  and SPED, SPEP and SPEF are significantly better than Baseline translation at  $p < 0.01$ . SPED and SPEP are also significantly better than SPEF at  $p < 0.01$ .

To cross check the results of the automatic scores with human evaluation, the following section reports the human evaluation results as to the overall quality of the four translations.

### 6.2.3.2 Human Evaluation Results

As with the preposition level evaluation, we employed two methods to find out the best translations. First of all, the number of No. 1 ranks for each translation assigned by at least three human evaluators was summarised and presented in Table 6.15.

Output ID	Top-ranked frequency	Proportion (out of 566)
Baseline	155	27.39%
SPEF	212	37.46%
SPEP	291	51.41%
SPED	319	56.36%

Table 6.15: Number and percent of times that each system was top-ranked (sentence level)

The percentages show the proportion of times out of the total number of sentences (566) that the system was evaluated to be the best translation.

The second method is to calculate the percentage of times that one system was ranked higher than any other system by at least three human evaluators (Table 6.16). The numbers in Table 6.16 indicate how often the column system was judged as better than the row system. Statistical tests show that the SPED is significantly better than SPEP at  $p<0.05$ . SPED, SPEP and SPEF are significantly better than the Baseline translation at  $p<0.01$ . SPED and SPEP are significantly better than SPEF at  $p<0.01$ . The results confirm the conclusion drawn from the automatic scores (see Table 6.14).

	Baseline	SPED	SPEP	SPEF
Baseline	/	48%	47%	44%
SPED	21%	/	10%	23%
SPEP	20%	14%	/	23%
SPEF	19%	37%	34%	/

Table 6.16: The percent of times that a system is judged as better than any other system (sentence level)

In summary, both automatic evaluation and human evaluation deemed that the order of the four systems from best to worst is SPED, SPEP, SPEF and Baseline.

Unlike at the preposition level, we did not conduct further qualitative analysis for each sentence since the focus of this study is translation of prepositions. However, further information is available in the work of Tatsumi and Sun (2008).

### 6.2.3.3 Correlation between Human and Automatic Evaluation at Sentence Level

The section above concludes that both human and automatic evaluation reached the same conclusion as to the overall ranking of the four systems. However, we have not checked their correlation level in terms of judging each sentence, i.e.

whether the higher scored translations were judged as better by human evaluators. Finding the automatic metric that correlates best with human evaluation at sentence level is of great importance (Russo-Lassner et al. 2005; Lin and Och 2004) as it could provide a detailed assessment of translation quality and pinpoint the problematic translations in a timely way without resorting to human evaluation. Our analysis in Section 6.2.1 reports that the inter-evaluator correlation is not satisfactory. Hence, in this section, we seek to find out the most suitable automatic metric for Chinese evaluation and to explore in what way automatic metrics can assist human evaluation.

To date, there is no definitive answer as to which is the best automatic metric for all languages. While most studies on correlation were tested on English as the target language, this study attempts to reveal which of the automatic evaluation metrics used in this study correlates best with human evaluation in terms of evaluation at sentence level for Chinese. As pointed out, while the four evaluators assigned ranks to each translation, the automatic metrics assigned discrete scores to each translation. A sample of the scores allocated is shown in Figure 6.5.

ID=1	H1	H2	H3	H4	BLEU	GTM1	GTM1.2	TER
Output 1	4	4	2	2	0.000	0.720	0.513	0.511
Output 2	2	2	2	1	0.335	0.833	0.562	0.417
Output 3	1	1	2	1	0.326	0.783	0.541	0.417
Output 4	3	3	2	2	0.000	0.750	0.532	0.500
ID=2	H1	H2	H3	H4	BLEU	GTM1	GTM1.2	TER
Output 1	4	4	2	2	0.495	0.927	0.513	0.255
Output 2	2	2	2	1	0.645	0.900	0.562	0.200
Output 3	1	1	2	1	0.702	0.950	0.541	0.151
Output 4	3	3	2	2	0.522	0.950	0.532	0.200

Figure 6.5: Sample of the human rankings and automatic evaluation scores

Using the approach of calculating Spearman's correlation between human and automatic evaluation for each sentence using the four subjects and then averaging

the sum by the total number of sentences has been questioned by Callison-Burch et al. (2008). The number of minimum subjects recommended in order to gain reliable statistics is different from one research field to the other. Kraemer and Thiemann (1987: 28) pointed out that between 10 and 20 subjects are commonly used in clinical tests in medicine while sociological studies rarely use fewer than several hundred subjects. Cohen et al. (2007) mentioned that in education studies, 30 is regarded as the minimum number of subjects in order to get valid statistical results. Obviously, statistical validity would be questionable with only four subjects per sentence.

An alternative way of checking the consistency between automatic evaluation metrics and human evaluation is to use Kappa statistics, as we did when checking the inter-evaluator correlation. We can transform the scores assigned by the automatic metrics into ranks first and treat each automatic metric as a special “evaluator”. Figure 6.6 below shows a sample after the transformation.

ID=2	H1	H2	H3	H4	BLEU	GTM1	GTM1.2	TER
Output 1	4	4	2	2	0.495	0.927	0.513	0.255
Output 2	2	2	2	1	0.645	0.900	0.562	0.200
Output 3	1	1	2	1	0.702	0.950	0.541	0.151
Output 4	3	3	2	2	0.522	0.950	0.532	0.200
ID=2	H1	H2	H3	H4	BLEU	GTM1	GTM1.2	TER
Output 1	4	4	2	2	4	2	4	3
Output 2	2	2	2	1	2	3	1	2
Output 3	1	1	2	1	1	1	2	1
Output 4	3	3	2	2	3	1	3	2

Figure 6.6: Sample of scores to ranks transformation

Following the steps we employed to calculate the inter-evaluator correlation, together with the Kappa formula and the statistical template, we obtained the following  $K$  values between each automatic metric and each human evaluator in Table 6.17.

	H1	H2	H3	H4	Overall
GTM (e=1)	0.427	0.569	-0.032	0.242	<b>0.302</b>
GTM (e=1.2)	0.397	0.479	-0.07	0.198	0.251
TER	0.415	0.513	-0.028	0.247	0.287
BLEU	0.347	0.397	0.058	0.296	0.275

Table 6.17: Kappa correlation between automatic and human evaluation at sentence level

We can see that GTM (e=1) correlates best with human evaluators based on the scores in Table 6.17. While in general, all automatic metrics correlate fairly or moderately with H1, H2 and H4, there is no positive correlation between automatic metrics and H3. Further examination of the results of H3 reveals that H3 assigned far more ties (i.e. two different translations were judged as equal in quality) than the others. Callison-Burch et al. (2008) pointed out that, unlike human evaluation, even when there is a slight difference between two translations automatic metrics generate two different scores. We calculated the number of tied pairs assigned by each evaluator and the percentage among all the translation pairs (2020) evaluated (Table 6.18).

	H1	H2	H3	H4
Number of Ties	125	26	1605	1029
Percentage	6.2%	1.3%	<b>79.5%</b>	50.9%

Table 6.18: Number of ties (and percentages) assigned by each evaluator

Apparently, the four evaluators have different standards in terms of assigning ties. H3 is far more likely to judge translations to be equal. This probably explains the low correlation between H3 and the automatic metrics and the low inter-evaluator correlation we presented in Section 6.2.1.

To minimise the effect of the ties assigned by humans, Callison-Burch et al. (2008) advocated a method for calculating the correlation between automatic metrics and human evaluation at sentence level. The results, including automatic

scores and human rankings were expanded into pair wise comparisons of any two systems. For each pair, we checked if the automatic scores were consistent with human ranking or not (that is the higher ranked system receives a higher score). The total number of consistent ranks was divided by the total number of comparisons to get a percentage. Pairs that human evaluators ranked as ties were excluded for the reason we just mentioned. The percentage of consistency between automatic metrics and human evaluators is reported in Table 6.19.

	H1	H2	H3	H4	Average
GTM (e=1)	61%	68%	71%	66%	<b>66%</b>
GTM (e=1.2)	58%	63%	68%	63%	63%
TER	58%	64%	70%	64%	64%
BLEU	51%	55%	65%	59%	56%

Table 6.19: Consistency level between automatic score and human ranking

Table 6.19 reveals that in this study, GTM (e=1) correlates best with the human evaluation at sentence level. Similar findings have been reported by Cahill (2009) in German evaluation which compared six metrics including the three metrics used in this paper. In addition, Agarwal and Lavie (2008) also mentioned that GTM and TER could produce more reliable sentence level scores than BLEU.

Having found that GTM (e=1) correlates best with human evaluation in this study, let us move back to a proposal we made at the beginning of the section, i.e. how to employ automatic metrics to assist human evaluation. As pointed out, the inter-evaluator agreements reported above are not very satisfactory and there are many ties assigned by human evaluators indicating either the translations are the same or human evaluators could not distinguish between them. Tied translations are not helpful in revealing the improvements and degradations in MT. Moreover, indistinguishable translations are the main source of discrepancy between the

human evaluators. Removing those indistinguishable translations can reduce the time required for human evaluation. We also believe it is a better approach to increase the inter-evaluator correlation than discarding evaluation results mentioned in Section 6.2.1.

A high correlation between automatic metric scores and human evaluation at sentence level indicates that if two automatic scores for a pair of translations are similar to each other, humans may also feel that the two translations are equal or indistinguishable. Therefore, it is useful to find out whether the difference reported by automatic scores reflects true qualitative differences in two translations and if there is a specific value for the difference between two automatic scores above which the majority of humans can also distinguish the two translations and agree with each other.

#### **6.2.4 Further Exploration**

In order to make use of automatic metrics to filter out those equal or indistinguishable translations, the data we gathered from the human evaluation were further explored to answer this question: How great a difference has to exist between two automatic scores so that the number of ties assigned by human translators is reduced to a minimum?

Generally speaking, we assume that the greater the differences between two automatic scores, the less likely that human evaluators would evaluate the two translations as ties. To check if this is true, we broke down the score difference into ten intervals (such as 0-0.1) and calculated the percentage of times that humans assigned ties for the pairs within each interval. For example, if the GTM ( $e=1$ ) scores for two translations are 0.64 and 0.53 respectively, the difference between these GTM scores (0.11) falls into the difference interval (0.1-0.2).



Figure 6.7 plots the the percentages of ties assigned by human evaluators to the translations within each score difference interval.

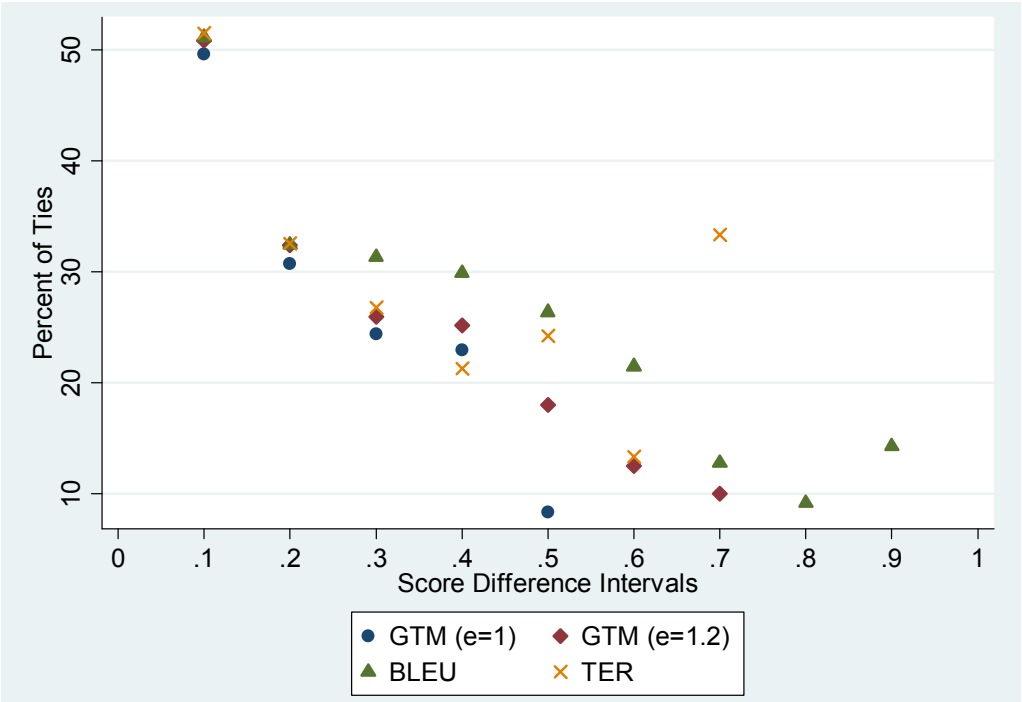


Figure 6.7: Percentage of ties assigned within each score interval

It can be seen that for any automatic metric, if the difference between two scores is not greater than 0.1, 50% or more translations will be evaluated as ties by human evaluators. Even when the difference between two scores is greater than 0.2, there are still around 30% of translations evaluated as ties. It can be confirmed that for TER and BLEU, even when two scores of two translations differ by more than 0.5, there is still a large percentage of translations that may be judged as ties by human evaluators.

The detailed percent of ties in each difference interval for each automatic metric is reported below in Table 6.20.

Score Difference Scales	GTM (e=1)	GTM (e=1.2)	TER	BLEU
[0,0.1]	49.63	50.89	51.50	51.17
(0.1,0.2]	30.77	32.43	32.56	32.47
(0.2,0.3]	24.41	25.98	26.77	31.36
(0.3,0.4]	22.95	25.21	21.28	29.91
(0.4,0.5]	8.33	18.03	24.22	26.37
(0.5,0.6]	/	12.5	13.33	21.48
(0.6,0.7]	/	10	33.33	12.80
(0.7,0.8]	/	/	/	9.17
(0.8,0.9]	/	/	/	14.29

Table 6.20: Percent of ties within each score difference interval

Based on the information displayed in Table 6.20 and Figure 6.7, we can conclude that GTM (e=1) would be the most suitable metric to be used in this study to pick up distinguishable pairs of translation for evaluation. This is also supported by the fact that our previous examination has already shown that GTM (e=1) correlates best with human evaluation at sentence level. In the analysis below, GTM refers to GTM (e=1) unless otherwise specified.

From Table 6.20 we can see that only when two GTM scores differ by more than 0.2, fewer than one third of the translations evaluated are ties, according to human evaluators. Therefore, if the focus of an evaluation is on the different translations from different systems, GTM scores at sentence level can be compared first and those pairs with their scores differing above the value of 0.2 can be selected and evaluated. The proposal can be stated as follows:

If an evaluation is to compare translations (for example, A and B) and the purpose is to reveal the improvements and degradations of A compared to B, in order to avoid a lot of ties from human evaluation which are of little value to the purpose of the evaluation, the GTM scores of the two translations at sentence level can be compared first, and then only those sentence pairs where the GTM scores of A and B differ by more than 0.2 should be retained for human evaluation.

Undoubtedly, to determine which score difference interval is the most suitable for a study depends on the purpose, the size, the time and the budget of the study. If

researchers/system developers want to compare the similarities between translations, then those sentences with a score difference below 0.1 can be selected for comparison. Alternatively, if one wants to focus on the benefits and drawbacks of one approach compared to a Baseline system, then translations with a score difference above 0.1 or (0.2) can be selected.

We did a post-hoc test on the effect of this method before applying it to our further experiments. As mentioned, all the sentences and translations were expanded into pair wise comparisons. We extracted translation pairs and their human rankings if the GTM scores of the two translations in a pair differed by more than (or were equal to) 0.2. We chose the upper bound (0.2) of the second interval  $((0.1, 0.2])$  because, while 30.77% of all translations are ties within this scale, most of these ties occur below the 0.2 value and just 6.6% of ties within this interval were assigned when two scores had a difference value equal to 0.2. Using 0.2 as the threshold, we can select more translations to be evaluated while keeping the number of possible ties to a minimum.

Using this filtering criterion, only 11.54% of all translations were left for evaluation. Since we already had their automatic scores and their human rankings, we can verify the validity of this filtering method by posing the following two questions. First, are the overall rankings of the four systems (Baseline, SPED, SPEP and SPEF) on the extracted pairs consistent with the conclusions we draw from the whole sample? Second, are the inter-evaluator correlations and the correlation between automatic and human evaluation improved?

As described above, we calculated the percentage of times that each system was evaluated as better than any other system. The numbers are reported in Table 6.21.

	Percentage of times
SPED	40.63%
SPEP	33.98%
SPEF	19.53%
Baseline	5.86%

Table 6.21: Percent of times that one system was evaluated as better than the others in the post-hoc test

The sequence of the four systems shown in Table 6.21 is consistent with our previous conclusion, i.e. SPED is the best and the Baseline is the worst. There is a smaller difference between SPED and SPEP compared to the differences between the other pairs. In other words, evaluating 11.54% of all translations leads to the same conclusion as evaluating all translations. The results verify the usability of our proposal with respect to saving evaluation time and cost.

The second aim of reducing the number of ties is to increase the inter-evaluator correlation to get more valid results. To test this, we calculated the inter-evaluator correlation of the four evaluators on the selected translation pairs. The four evaluators enjoyed high agreement. 66.07% of all pairs received a majority vote. The Kappa value for the inter-evaluator agreement is 0.336 compared to 0.273 obtained from all translations. Although both  $K$  values belong to the category of Fair agreement, there is indeed noticeable improvement. The advantage is that no evaluation result has to be discarded in order to increase the correlation level.

Finally, we also checked the correlation between automatic and human evaluation. Since we used GTM ( $e=1$ ) to filter out sentences, we can only check the consistency level between GTM and human evaluation on the extracted pairs of translations. Following the same procedure described in Section 6.2.3.3, we checked the number of sentences for which the automatic and human evaluation

agreed with each other (Table 6.22). This consistency level is slightly higher than that reported in Table 6.19 (66%).

	H1	H2	H3	H4	Average
GTM (e=1)	79.08%	90.56%	32.91%	67.65%	<b>67.55%</b>

Table 6.22: Consistency level between automatic score and human evaluation in the post-hoc test

In summary, using our proposed criteria to filter the number of sentences to be evaluated by humans not only can save time and resources but also improves validity. This proposal can be easily applied to other studies. However, as mentioned, the specific values have to be dependent on the purpose of the study and the available resources.

### 6.3 Summary

We come to the conclusion from our analysis that the unmodified general SPE system can produce better translations both at sentence level and preposition level than the Baseline RBMT systems. A modification to the phrase table of the SPE system failed to outperform the unmodified SPE system but generated significantly better translations than the Baseline system, especially on sentence level translation. The most frequently corrected preposition error by the general SPE system was Incorrect Position. Incomplete translation of preposition, especially *in* or *on* was the second most frequently corrected error. One of the modified SPE modules did bring some unique improvements. The main difference between SPED and SPEP and SPEF is the size of their phrase table. Results show that the more phrases (i.e. the more information that is presented in the phrase table) in an SPE module, the better the translation of prepositions. In other words, the translation of prepositions is not in isolation but closely related with

translation of other parts of a sentence. Therefore, in our further research, we proposed general approaches instead of ones that are preposition focused.

We also found out that one major factor influencing the correlation between human evaluators is the indistinguishable translations. Some studies propose discarding results of the evaluators that correlates worst with the others in order to obtain reliable results. However, this approach is not suitable for experiments with a limited number of evaluators. Moreover, the approach does not help in saving time and cost. Therefore, we propose to make the evaluation purpose-specific and reduce the number of evaluations so that the evaluation task could be simplified. This approach is advantageous in the following ways. First, it could save time and resources. Second, by simplifying the evaluation process, the reliability of the results could be enhanced instead of discarding data. Finally, it could help system developers to determine whether an improvement in an automatic score is significant or not or whether/how human evaluation should be conducted.

Another important by-product of the current study is that in terms of Chinese IT document evaluation, all the three automatic metrics involved in this study correlated well with human evaluation at system level. However, at sentence level, GTM ( $e=1$ ) stands out as the best automatic metric.

## **Chapter 7: Dictionary Customisation and Source Pre-Processing**

Human post-editing plays a decisive role in obtaining the final translation of publishable quality in the localisation industry. Post-editing, therefore, has attracted some attention and inspired much research. On the other hand, pre-processing has been relatively less studied, especially for proprietary types of RBMT systems such as the system used in this study. Pre-processing is defined in previous chapters as the first step in a translation process which includes any preparation that facilitates an MT system in analysing the input more effectively, and consequently, generating a better translation. Another approach serves for the same purpose is dictionary customisation which is a way of tuning the MT system in order to get better translation. Similar to the pre-processing approach, dictionary customisation is conducted before inputting a source text into an RBMT system; hence, dictionary customisation and source pre-processing are introduced within the same chapter. We first introduce how to add a supplementary dictionary of prepositions and prepositional phrases, which was generated automatically by the use of an SMT system, to the RBMT system. In Section 7.1.1, we outline the design of this proposal and the set-up of the project. We then carry out an evaluation and report the results (7.1.2). The second approach is a general pre-processing principle which aims to change the source English into a more target-language-similar or RBMT-system-friendly language. The rationale of proposing this approach is illustrated in Section 7.2.1. Its implementation and evaluation results are presented in Section 7.2.2 and 7.2.3 respectively.

## 7.1 Automated Preposition Dictionary Extraction

In our pilot project in Chapter 4, we showed that adding the main user dictionaries created by Symantec produced significantly better Chinese translation than the output without user dictionaries. However, there were few entries concerning prepositions in the main UD; hence, few improvements of translation in the prepositions were observed (see Section 4.5.1). Following the traditional process of encoding entries to build a preposition dictionary was not desirable. Arnold et al. (1994) pointed out that user dictionaries were the most expensive MT component to construct both in terms of cost and time. The task would entail employing several expert linguists and an entry has to be tested in several contexts to avoid unnecessary degradation before it is added to a dictionary.

Dugast et al. (2009) reported part of their ongoing research on quickly obtaining an extra phrasal dictionary for their RBMT system through making use of the bilingual phrases generated from training an SMT system. They obtained 5,000 and 170,000 entries from two corpora respectively. Each entry was then validated individually by checking if adding the entry could bring higher BLEU scores or not. The final test results showed that adding the extra entries increased the BLEU score by 7% over the Baseline system (from 0.2488 to 0.2665).

As they mentioned, this was just part of the ongoing work. Several problems exist in the above experiment. First of all, to validate the efficiency of each entry through BLEU is questionable due to the low correlation between BLEU scores and human evaluation, especially at sentence level. In addition, no detailed linguistic comparison as to what changes were brought by the extra dictionary was conducted.



Based on the work of Dugast et al. (2009), we designed a process to extract a supplementary preposition dictionary for the RBMT system automatically. As the focus of the current study is on translation of prepositions, we only included entries containing prepositions. By restricting the entries for a supplementary prepositional phrase dictionary, changes in translation of prepositions introduced by the new entries can be easily pinpointed.

### 7.1.1 Experiment Set-up

We defined four basic steps to extract entries with prepositions from the bilingual parallel corpora. The extraction procedure of this preliminary study is illustrated in Figure 7.1.

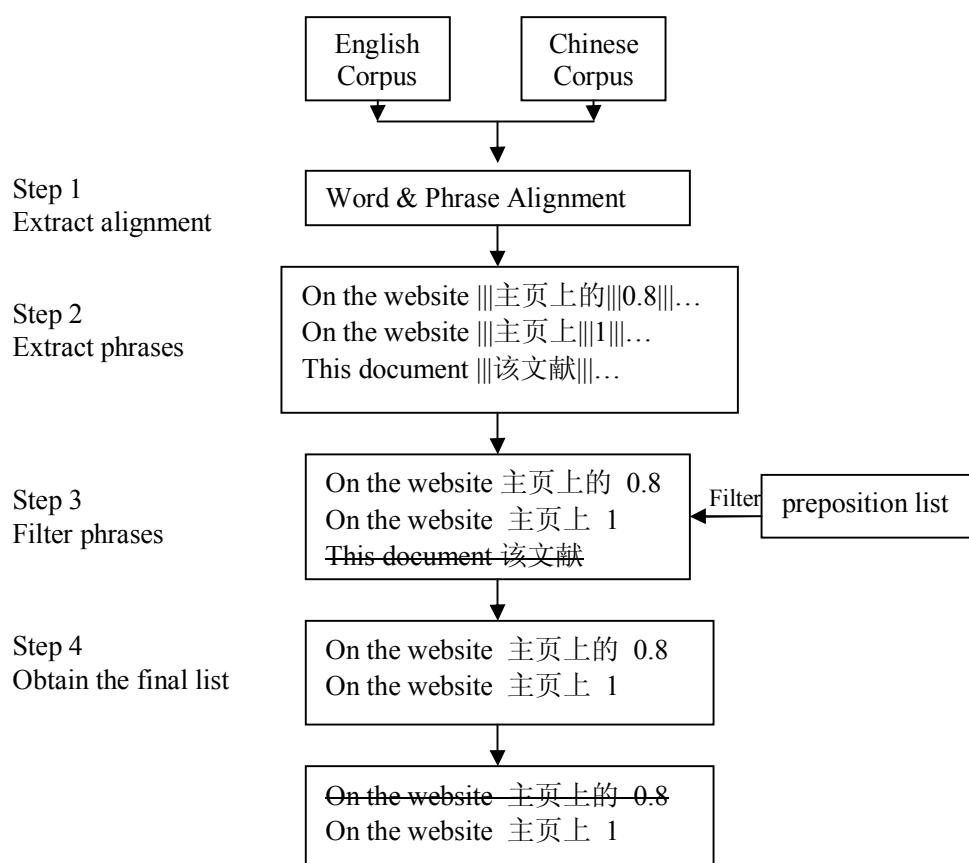


Figure 7.1: Pipeline for extracting prepositional entries

Let us go through each step in full detail. The bilingual corpora involved are the preposition corpora constructed at the beginning of the study. There are 5951 sentences each of which contains at least one preposition on the English corpus side. The Chinese corpus is the reference translation of the source English sentences.

- Step 1 Obtain the word/phrase alignment between the bilingual corpora

Using the Moses toolkits and following the tutorial on its website, we extract the word alignment information between the bilingual corpora. The process is similar to training an SMT system but the purpose is to generate the alignment information and the phrase table.

- Step 2 Extract phrases

The phrase table generated above contains much information such as the word alignment information and the translation probabilities of two phrases. The meaning of the phrase table has been explained in detail in the Chapter 2 (see Section 2.2.3). In this step, we only extract the bilingual phrases we need: a source word or phrase followed by the corresponding target language translation and their translation probability.

- Step 3 Extract the bilingual phrases with prepositions

The result of step 2 is a table with bilingual phrase pairs. We then filtered the phrase table using the pre-defined preposition list (the list of the ten prepositions defined in Chapter 4). If a preposition was present on the English side of a bilingual phrase pair, then this phrase pair was kept as an entry in the dictionary,

otherwise the phrase was removed. Therefore, the first entry in Step 3 of Figure 7.1 was retained while the third entry was removed.

- Step 4 Select the bilingual phrases with the highest probability

From the last step we had a list of bilingual phrase pairs each containing one or more prepositions. One problem associated with the list is that there are repetitive entries. For some English entries, there are multiple translations; or, multiple English entries correspond to the same Chinese translation. The difference between these entries lies in the probabilities generated during the training process. To ensure the quality of the dictionary and to reduce the noise to the minimum, only entries with the highest probabilities were encoded in the dictionary.

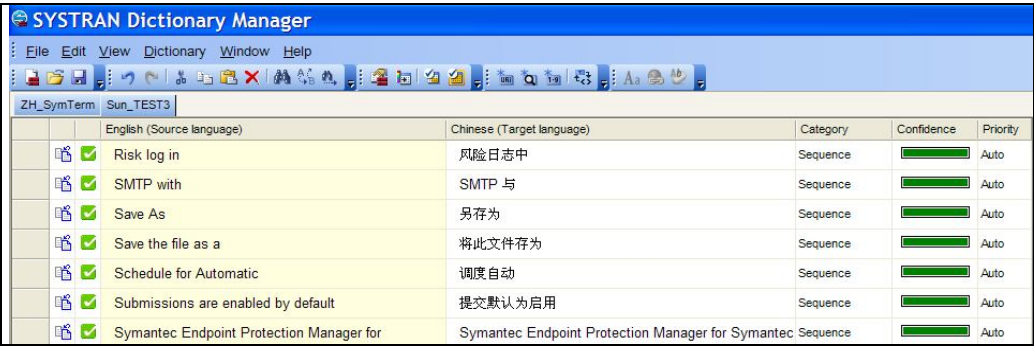
The final list contained 811 entries, each of which contains at least one of the ten prepositions. The number of phrases containing one of the ten specific English prepositions is plotted in Table 7.1.

in	to	of	on	for	from	with	by	about	as
72	328	34	53	101	73	98	19	66	16

Table 7.1: Number of entries of each preposition extracted from the phrase table

We built a special preposition user dictionary using these entries following Systran’s user manual on how to encode an entry into a dictionary. In addition, Symantec has created a script and an interface that can transform a batch of words into a dictionary format automatically. It is necessary to remind the reader that the phrases extracted are not “phrases” in the traditional linguistic sense. Instead, they are just chunks of words. Hence, these “phrases” are more similar to the meaning of “sequence” defined by Systran as “those words and phrases (especially fixed expressions) that do not undergo linguistic analysis, but that are accepted ‘as-is’

for the final translation.” Therefore, all entries were encoded as a “sequence” in our preposition dictionary. Figure 7.2 is a screenshot of the final dictionary.



	English (Source language)	Chinese (Target language)	Category	Confidence	Priority
	Risk log in	风险日志中	Sequence		Auto
	SMTP with	SMTP 与	Sequence		Auto
	Save As	另存为	Sequence		Auto
	Save the file as a	将此文件存为	Sequence		Auto
	Schedule for Automatic	调度自动	Sequence		Auto
	Submissions are enabled by default	提交默认为启用	Sequence		Auto
	Symantec Endpoint Protection Manager for	Symantec Endpoint Protection Manager for Symantec	Sequence		Auto

Figure 7.2: Automated extracted preposition dictionary

### 7.1.2 Translation Evaluation

We first translated the test sample (1000 prepositions) using Systran. We then translated the sample again, this time adding the supplementary preposition dictionary. Recall that the main purpose here is to check if adding our automated extracted preposition dictionary can bring additional gains in terms of translation of prepositions compared to the output from the existing dictionaries which had few entries relating to prepositions.

Two sets of translations were generated. The first translation was the Baseline translation and the second translation was named as Exdic\_output. As in the experiments explained in previous chapters, GTM (e=1), GTM (e=1.2), BLEU and TER were employed to measure the translation quality. Table 7.2 reports the scores of the translations without and with the supplementary preposition dictionary.

Output ID	GTM (e=1)	GTM (e=1.2)	BLEU	TER
Baseline (without the supplementary dictionary)	0.417	0.357	0.241	0.538
Exdic_output (with the supplementary dictionary)	0.426	0.360	0.243	0.537

Table 7.2: Automatic evaluation scores of the translations with/without the supplementary preposition dictionary

All the automatic scores exhibit the same trend, i.e. translation with the supplementary dictionary is better than the Baseline translation; however, the difference between the scores is small. Statistical significance test shows that there is no significant difference between the automatic scores of the translations with and without the new supplementary dictionary. Since all four automatic metrics report no significant difference between the translations, instead of conducting a large scale human evaluation, the author herself conducted a preliminary comparison in order to reveal what changes were brought by the supplementary dictionary.

Most prepositions are translated more or less similarly with the Baseline translation with only 75 different translations after adding the supplementary dictionary. We compared the 75 translations (Exdic\_output) with the corresponding Baseline translation to check which translation was better. It was found that out of these 75 different translations between the new output and the Baseline, 20 sentences were evaluated as equal, 30 translations with the supplementary dictionary were evaluated as better than the Baseline and for the remaining 25 sentences, Baseline was judged as better. In other words, there is almost an equal number of improvements and degradations brought about by the preposition supplementary dictionary.

Another finding from the analysis of the 75 different translations is that all of the changes are related to preposition translations with only one exception. By restricting the entries in the dictionary, i.e. only translation of prepositions were encoded, we can easily pinpoint the changes to the translation of prepositions.

Since there is no significant difference in terms of the quality of the translation with and without the supplementary dictionary, we did not conduct an in-depth comparison as to what errors could be corrected. Instead, the author briefly compared the two translations and examined the types of changes brought by the supplementary dictionary. The differences were divided into two groups: difference of lexical translation and difference of word order.

These two groups were the prototype error categories summarised from the Symantec internal user report in Chapter 3. The error typology of prepositions constructed in Chapter 3 can also be described by these two groups. For example, incorrect-position errors belong to the group of word order. And the remaining errors, incorrect lexical selection, incomplete translation and translation missing can be generally regarded as change of lexical translation. In summary, these two groups can broadly represent most of the differences between translations.

In this study, a change is considered as a lexical change if only translations for certain words or phrases change without changing the structure of the sentence. Changes in punctuation marks are also included in this group. A change is considered as a word order change if there is a difference in terms of the word order, including word level order change and long distance phrase level order changes. The author first categorised whether the difference between the two translations belongs to the category of lexical translation or a change in word

order. There is almost the same number of changes in the two categories with 42 lexical changes and 37 word order changes.<sup>51</sup>

We also found that both improvements and degradations were generated by the supplementary preposition dictionary in each of the two groups. Example 7.1 shows lexical changes brought about by the supplementary preposition dictionary. 7.1a is an example of improvement generated by the supplementary dictionary and 7.1b shows a degradation generated by the preposition dictionary.

---

Example 7.1a - Improvement

**Source:** Do not **log on to** the Symantec Policy Manager.

**Baseline:** 不要 注册 到 Symantec Policy Manager 。 /pīnyīn: zhù cè dào/

**Gloss:** Do not register to Symantec Policy Manager.

**Exdic\_output:** 不要 登录 Symantec Policy Manager 。 /pīnyīn: dēng lù/

**Gloss:** Do not log on Symantec Policy Manager.

---

Example 7.1b - Degradation

**Source:** **For** any update, you can select whether the update occurs within minutes of the scheduled time.

**Baseline:** 对于 所有 更新, 您 能 选择 这次 更新 是否...。 /pīnyīn: duì yú/

**Gloss:** For all updates...

**Exdic\_output:** 为 任何 更新, 您 能 选择 这次 更新 是否...。 /pīnyīn: wéi/

**Gloss:** For any updates...

---

In 7.1a, the Baseline translation is not as good as the new translation. The main problem is that in the Baseline the Chinese translation corresponds to “*register*” rather than to “*log on*”. Besides, the whole phrase *log on to* is more often translated into one Chinese word (two characters) and omits translating *to* as in the Exdic\_output. With regard to the second example (7.1b), the Exdic\_output is

---

<sup>51</sup> Note that in a few cases the two translations differ both in lexical translation and word order. Hence, the total number of changes is not equal to 75.

worse than the Baseline in terms of the translation of the preposition *for*. The Exdic\_output translation in 7.1b is more often used in contexts meaning “*for the purpose of*” instead of the context in this example.

Examples in 7.2 below show some word order changes brought about by the supplementary preposition dictionary. Again, 7.2a and 7.2b present an improvement and degradation respectively. In both examples, the differences between the orders of the two translations can be seen from the different positions of the highlighted words. Following the practice in Chapter 3, special function words of Chinese are spelled out and their functions are noted in brackets.

---

#### Example 7.2a - Improvement

**Source:** In Test mode, you can **apply** the policy **to devices** and generate a Client Control Log.

**Baseline:** 在 测试 方式 ， 您 能 **应用** 这项 策略 **到 设备** ， 并且 引起 客户端 请 控制 日志 。

**Gloss:** In test mode, you can apply the policy to devices...

**Exdic\_output:** 在 测试 方式 ， 您 能 将 策略 **应用 于 设备** ， 并且 引起 客户端 请 控制 日志 。

**Gloss:** In test mode, you can JIANG (active marker) the policy apply to devices...

---

#### Example 7.2b - Degradation

**Source:** To **exclude** a file **from scans**.

**Baseline:** **从 扫描 排除** 文件。

**Gloss:** From scan exclude file.

**Exdic\_output:** **排除** 将 文件 **从 扫描**。

**Gloss:** Exclude JIANG (active marker) file from scan.

---



Literal word-for-word translation of the English sentence in 7.2a results in an extremely awkward Chinese sentence (as in the Baseline). Chinese expresses voice or tenses through functional words and word order instead of morphological inflections. The Exdic\_output in 7.2a successfully added the function word “JIANG” and generated a translation with correct word order. However, for example 7.2b, while the Baseline system generated a correct active voice translation, in the Exdic\_output translation the function word was inserted at a totally incorrect position resulting in an incomprehensible translation.

Within each of the two groups, the number of better, worse and equal translations with the supplementary dictionary compared to the Baseline translation was calculated. The purpose was to reveal if the supplementary dictionary could bring better lexical translations or better word order. Figure 7.3 plots the number of translations that were evaluated as better in each group. From Figure 7.3 we could see that in terms of lexical selection, the supplementary dictionary is better than the Baseline translation; however, with regard to word order, the Baseline translations are better than the ones with the supplementary dictionary.

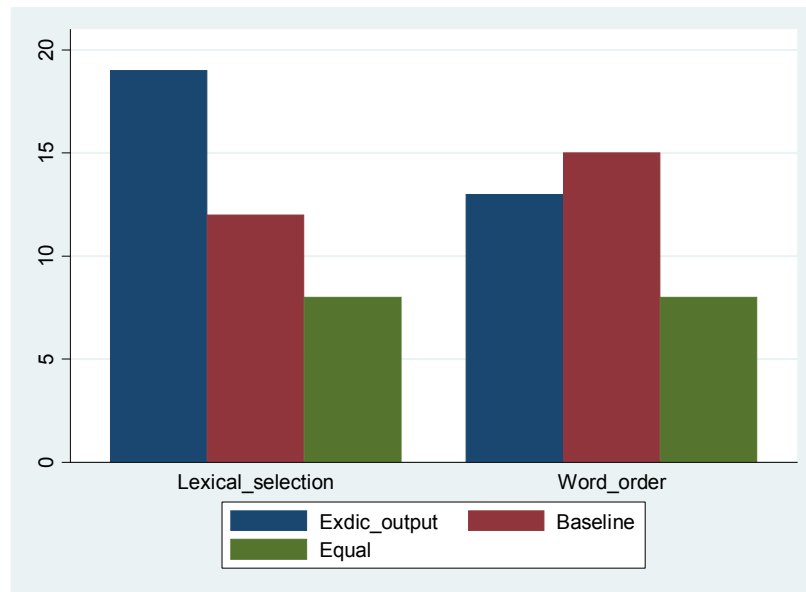


Figure 7.3: Improvements, degradations and equal translations

It is not surprising that the lexical translation with the supplementary dictionary is better. The main reason for the degradation in word order in some cases is probably that since all entries were encoded as sequences, the translations were “protected” and would not go through the regular analysis. In addition, most of the phrases encoded were long sequences ranging from 2 words to 7 words. Once a phrase was found in the dictionary, the corresponding translation encoded in the dictionary could be inserted; however, the inserted translation might interfere with the internal analysis rules of the RBMT systems leading to incorrect word order generation. Overall, there is no significant difference between the two translations.

To sum up, using the Moses statistical toolkits, a tailor-made preposition dictionary was automatically extracted. Both the automatic evaluation and the preliminary evaluation conducted by the author indicated that this approach could be beneficial. The advantage of this preposition dictionary is that it could be obtained automatically without manually crafted rules. In addition, it contains

entries only related to prepositions; hence, changes could be easily detected. However, this dictionary does not bring about significant improvement. The preliminary analysis revealed that this automated dictionary could bring some better lexical translations but may affect word order. To reveal the major reason for the degradations could be an interesting point for future research.

Word order is a big challenge faced by any MT system due to the grammatical difference between languages, especially between English and Chinese. While compiling a dictionary may not be very effective, changing the structure of an English text into Chinese grammar has attracted much attention recently. Some related work is introduced in the next section, and based on previous work, a more general pre-processing approach for English to Chinese translation is proposed.

## **7.2 Statistical Source Reconstruction**

Recently, a new pre-processing approach that suggests changing the source text to be closer to the structure of the target language has been reported. Wang et al. (2007) reported that transforming Chinese sentences, by using hand-coded linguistic rules to be closer to English in terms of syntactic structure, could increase the scores of the final translation by an MT System. Xu and Seneff (2008) transformed English texts into "Zhonglish" (English words in Chinese structure) before translating them by an MT system and found that human evaluations favour the translation output from "Zhonglish" as the source to the translation of the original English texts. A number of other researchers have also described their pre-processing methods on other language pairs. Xia and McCord (2004) reported the effect of automatically learnt rewriting patterns in improving English and French translation. Crego and Marino (2006) proposed an approach to coupling reordering and decoding in SMT and reported significant improvements in

translation quality between English and Spanish. These papers focused on incorporating syntactic information into SMT systems with rules either hand-crafted or automatically extracted. Babych et al. (2009) conducted a similar study for an RBMT system. They applied "construction-level human evaluation" to discover systematically mistranslated structures and then to "create automatic pre-editing rules to make the constructions more tractable for an RBMT system" (2009: 36). Their study only concentrated on some of the most frequently occurring light verb constructions ("verb phrases with a semantically depleted verb and its objects, such as take part"). In addition, they still needed to compose the pre-editing rules manually.

In this section, we introduce a new statistical pre-processing model for the RBMT system. The design of the current model differs from the previous ones in the following ways: firstly, the pre-processing model is designed for an RBMT system while most of the previous work focuses on SMT systems; secondly, the transformation process is automated without any hand-coded rules; thirdly, the translation direction is from English to Chinese which is less studied compared to Chinese to English translation.

The remainder of the section is organised as follows. In Section 7.2.1, we explain the rationale of our pre-processing model followed by a pilot test and the linguistic analysis of the new test. Section 7.2.2 presents the general methodology of the pre-processing model and the experimental set-up. Finally, some evaluation results are reported in Section 7.2.3.

### 7.2.1 Rationale

Our method was inspired by a test related to “round-trip translation” (Somers 2005), one intuitive evaluation approach usually (and especially) used mainly by monolingual lay people to determine the quality of an MT system. “Round-trip translation” includes translating a text in one language into a second language (Forward-Translation) and then translating it back into the original language (Back-Translation). In cases where the evaluators do not know the target language or no target language reference is available, “round-trip” translation seems to be an intuitive and easy solution for judging the performance of an MT system based on the assumption that the Back-Translation can represent the quality of the Forward-Translation. However, some considered it as a rather naïve or inappropriate way of measuring translation quality. For example, by comparing the BLEU scores of the Forward-Translation and the Back-Translation, Somers (2005) claimed that overall “round-trip translation” was not suitable for MT evaluation as Back-translations tended to get higher scores than Forward-translations. Others claimed that it could be useful at sentence level evaluation (Rapp 2009).

Whether “round-trip” translation could or could not be used as a means of MT evaluation is not the focus of this study. Forward-Translation and Back-Translation are defined differently in the current study. Generally speaking, Forward- and Back-Translation occur across two different languages, with Forward-Translation into the target language and Back-Translation into the source language. In our study, the “Forward-” and “Back-Translations” are in the *same* language (both Chinese). To avoid confusion with the traditional definition of Forward-Translation and Back-Translation, a new set of symbols are used.

Procedure 1, illustrated below, explains how to obtain this new pair of translations for comparison.

<p>Procedure 1: Steps to Obtain New Forward-Translation and Back-Translation</p> <ol style="list-style-type: none"> <li>1) Input a source English text (<math>E_o</math>, <b>Original English</b>) into an RBMT system and get the target language translation (in this study, Chinese). Name this translation as <math>ZH_{MT}^F</math> (it can be regarded as a “Forward-Translation” from English to <b>Chinese</b> by the <b>MT</b> system)</li> <li>2) Input the Chinese reference <math>ZH</math> (which is human translation or human post-edited MT translation of the above English text) into the same RBMT system and get an English translation output. Name this English translation as <math>E_{MT}</math> (it can be regarded as a “Forward-Translation” from Chinese to <b>English</b> by the <b>MT</b> system)</li> <li>3) Input <math>E_{MT}</math> from the above step into the same RBMT system and get the final Chinese translation output. Name this Chinese translation as <math>ZH_{MT}^B</math> (it can be regarded as a “Back-Translation” of the <b>Chinese</b> reference mentioned in the second step. The whole process is <math>ZH \rightarrow E_{MT} \rightarrow ZH_{MT}^B</math> (translate the Chinese reference into English and then translate back into Chinese by the MT system))</li> </ol>
--

To see which translation is better, the Chinese “Back-Translation” ( $ZH_{MT}^B$  from step 3 in Procedure 1) or the Chinese “Forward-Translation” ( $ZH_{MT}^F$  from step 1 in Procedure 1), two samples were randomly selected from the corpus used in the dictionary test. The Chinese reference translations were extracted from Symantec’s in-house Translation Memory. Table 7.3 gives the sizes of the two samples.

Corpus (#Sentences)	# English Words	# Chinese Words
Sample 1 (500)	9,830	10,703
Sample 2 (1,000)	15,915	17,257

Table 7.3: Pilot samples for the comparison of back- and forward-translation

Samples 1 and 2 were processed according to the three steps in Procedure 1 and two pairs of Chinese translations ( $ZH_{MT}^F$  and  $ZH_{MT}^B$ ) were generated. As in the

previous chapters, the three automatic metrics were employed to obtain the overall scores of the translations by comparing them to the Chinese reference. Table 7.4 reports the scores of  $ZH_{MT}^F$  and  $ZH_{MT}^B$  of the two samples.

	Sample 1	Sample 2	Sample 1	Sample 2	Sample 1	Sample 2	Sample 1	Sample 2
	GTM (e=1)		GTM (e=1.2)		BLEU		TER	
$ZH_{MT}^F$	0.426	0.422	0.352	0.316	0.238	0.184	0.542	0.594
$ZH_{MT}^B$	0.529	0.511	0.428	0.394	0.357	0.294	0.402	0.464

Table 7.4: Automatic evaluation scores of the back- and forward-translations

Statistical tests of the scores in Table 7.4 confirm that Chinese “Back-Translation” ( $ZH_{MT}^B$ ) is significantly better (at  $p<0.01$ ) than Chinese “Forward-Translation” ( $ZH_{MT}^F$ ) in terms of all the automatic scores for both samples. The next section compares the two translations in detail and reveals one key reason for their differences. Finally, a new pre-processing model is proposed based on that key reason.

### 7.2.1.1 Qualitative Comparisons

One possible reason for the differences between the scores of  $ZH_{MT}^F$  and  $ZH_{MT}^B$  relates to what Somers (2005) mentioned about the difference between Forward-Translation and Back-Translation in his tests:

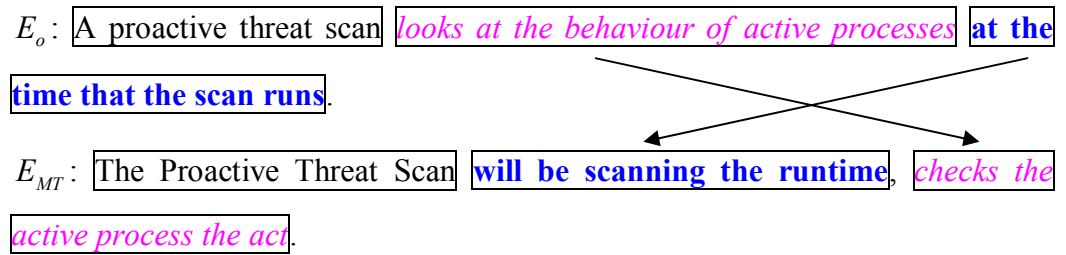
Although systems perform source-text analysis to a certain extent, when all else fails they resort to word-for-word translation, and where there is a choice of target word they would go for the most general translation. Clearly, when the input to the process is difficult to analyse, the word-for-word translation will deliver pretty much the same words in the BT [Back-Translation] as featured in the original text (2005: 30).

Hence, in our test when the Chinese reference sentences were translated into English ( $E_{MT}$  in Procedure 1) due to some failed source analysis the system generated some word-for-word translation in English with some Chinese

flavoured structures. When this English translation was translated back to Chinese, a second round of word-for-word translation generated some translations that were the same as the original Chinese reference sentences. In other words, one assumption about  $E_{MT}$  is that it contains target-language friendly or at least MT-friendly structures and that is why its translation ( $ZH_{MT}^B$ ) is better than the translation of the source English text ( $ZH_{MT}^F$ ). This assumption arose after comparing the  $E_{MT}$  and the  $E_o$  text. The following example (example 7.3) shows their differences.

---

Example 7.3




---

We will just focus on the two major differences between these two English sentences marked by bold font and italics. In English, the adverbial phrase (in this example, “**at the time that the scan runs**”) is placed after the verb of the sentence (“looks at”) while in Chinese, it is usually placed before the verb. The  $E_{MT}$  sentence shows this characteristic by moving the phrase (which is “**will be scanning the runtime**” in  $E_{MT}$ ) in front of the verb (“checks”). Another difference is the position of the modifier and the modified nouns. In the source English sentence, the modifiers follow the modified nouns in an attributive clause (such as “**the time that the scan runs**”) or in prepositional phrases (such as “*the behaviour of active processes*”). However, in Chinese, the modifiers appear before the modified nouns. Again, the  $E_{MT}$  sentence exhibits this grammatical



characteristic: “**the time that the scan runs**” was changed to “**scanning the runtime**” and the prepositional phrase “*the behaviour of active processes*” was changed to “*the active process the act*”, both of which put the original modifier before the modified nouns. The differences between the source English sentence ( $E_o$ ) and the English sentence  $E_{MT}$  are most likely the reason why the  $ZH_{MT}^B$  gets higher automatic scores than  $ZH_{MT}^F$ .

From the above analysis and the improved automatic evaluation scores, one hypothesis which can be derived is that if an English source sample can be pre-processed into the structures of  $E_{MT}$ , its Chinese translation could be better than the direct Chinese translation of this English sample. Therefore, in the next section, we introduce a statistical model to automatically pre-process the source texts to an RBMT system into the structures similar to that of  $E_{MT}$ . As we mentioned in the introduction, there are already prior studies showing that changing a source text to be closer to the target language could improve the translation output (Wang et al. 2007; Xu and Seneff 2008).

### 7.2.2 Experiment Set-up

To test the hypothesis that “If we pre-process an English sample into the structure of  $E_{MT}$ , the final translation should be better”, we need a model that can learn the structures of  $E_{MT}$  and automatically transform a new English sample into similar structures. An SMT system, which is trained using two parallel corpora (a source language corpus and a target language corpus) and some statistical methods to try to generate the best target translation for a source sentence, is a good candidate to conduct this transformation. SMT systems have been applied to post-edit the output of RBMT systems (which is known as SPE). For example, in Chapter 6 we

reported the effectiveness of an SPE system in improving the RBMT output quality. Let us remind the readers of the basic processes of an SPE system: first, a corpus is translated using an RBMT system from one language (let us continue with the example of English) into a target language (Chinese). Secondly, an SMT system is trained using this Chinese translation as the “source language” and the Chinese reference translation as the “target language”. The SMT system will learn how to post-edit raw Chinese RBMT output into the corresponding Chinese reference translation. Thirdly, once a new English text is translated using the same RBMT system into Chinese, the translation can be input into the trained SMT system to be post-edited into a revised translation.

Our proposal combines an SMT system and an RBMT system in a similar but a novel manner, i.e. using an SMT system to pre-process the source to the RBMT system instead of post-editing the output from the RBMT system. The process is described in Procedure 2 below:

### Procedure 2: Statistical Pre-Processing

- 1) Input a Chinese reference corpus into an RBMT system and get the English translation output. As in Procedure 1, name it  $E_{MT}$  (an **E**nglish translation of a Chinese corpus from the **R**BMT system). It will function as a “pivot” English with some Chinese characteristics or the RBMT-system friendly structures;
- 2) Train an SMT system using the  $E_{MT}$  corpus as the “target” text and the source English corpus  $E_o$  as the “source” text. Let the SMT system learn how to translate or pre-process the source English into  $E_{MT}$  style English (a kind of pseudo English);
- 3) Input a new English sample (with no sentences that have appeared in the training corpus) into the trained SMT system. The output will be an English text with  $E_{MT}$  style or flavour.
- 4) Translate the pre-processed sample from the last step into Chinese using the RBMT system. The final output is **P**re-**p**rocessed **C**hinese **M**achine **T**ranslation output ( $ZH_{MT}^{PP}$  for short).

As in the SPE experiment, the SMT system used was Moses and the Baseline MT system was Systran. The same test sample and training corpus used in Section 7.1 were used. Recall that the training corpus and the test sample are special preposition corpora in the way that each sentence contains at least one of the top ten frequent prepositions (see Section 4.3.2). Let us review the sizes of the training and the test sample (Table 7.5). Since this training corpus was selected under the same criteria as the test sample, it is named as In\_Domain corpus.

	# Sentences	# Words
Training set (In domain)	5,951	84,349
Test set	944	15,916

Table 7.5: In-domain training corpus and test set

As mentioned before, this is a comparatively small corpus in terms of training SMT systems. Although our experiment on SPE in Chapter 6 has proven that even

using this small corpus can get significantly better translations, we are not sure if this is enough for this pre-processing experiment. Since we had at our disposal the translation memory from which the corpus was extracted, besides the in-domain corpus, we also used the whole TM. This TM is a mixed corpus in which sentences may or may not contain prepositions. We named this corpus the mixed-domain corpus. Three training corpora were randomly selected from the TM. The first corpus contained the same number of sentences as the in-domain corpus (Mixed\_small). The second and third corpora were larger randomly extracted corpora (Mixed\_medium and Mixed\_large). The purpose of using another two larger corpora was to see if the size of the training corpus would affect the performance of the pre-processing model. In addition, comparing in-domain to mixed-domain corpora could reveal the importance of training data on the final output. The three additional training corpora are listed in Table 7.6.

	# Sentences	# Words
Mixed_small	5,951	55,846
Mixed_medium	9,934	106,457
Mixed_large	94,622	1,250,905

Table 7.6: Three mixed-domain training corpora

For each of the training corpora listed in Table 7.5 and Table 7.6, the four steps explained in Procedure 2 were repeated to get the final Chinese translations. Four different translations of the test sample were generated from four pre-processing models trained using the four different training corpora. The Baseline translation was obtained by translating the test sample using Systran without any other pre-processing. The final five translations, namely, the Baseline translation (Baseline), the translations of the three random training corpora respectively ( $MT_{mixed}^{small}$  from Mixed\_small,  $MT_{mixed}^{medium}$  from Mixed\_medium and  $MT_{mixed}^{large}$  from

Mixed\_large) and the translation using the in-domain training corpus ( $MT_{in-domain}$  from the In\_domain corpus) were scored by comparing them to the reference translation using the three automatic evaluation metrics. The next section reports the scores of these translations and gives a brief analysis of the translation results.

## 7.2.3 Results

### 7.2.3.1 Automatic Evaluation Results

Table 7.7 below reports the automatic scores of the Baseline from the original source and the four translations from pre-processed sources by the pre-processing models.

	GTM(e=1)	GTM (e=1.2)	BLEU	TER
Baseline	0.417	0.357	0.241	0.538
$MT_{mixed}^{small}$	0.420	0.352	0.223	0.550
$MT_{mixed}^{medium}$	0.428	0.360	0.230	0.544
$MT_{mixed}^{large}$	0.457**	0.384**	0.275**	0.506**
$MT_{in-domain}$	0.445**	0.374**	0.265**	0.526

Table 7.7: Automatic evaluation scores of pre-processed translations

From the scores above we can see that  $MT_{mixed}^{large}$  is the best translation, followed by  $MT_{in-domain}$ . The small and medium pre-processing modules failed to outperform the Baseline system. We performed significance tests on the improvements of the automatic scores compared to the Baseline translation using bootstrapping re-sampling (Koehn 2004). Scores with \*\* indicate the translation is significantly better than the Baseline translation at  $p < 0.01$ .  $MT_{mixed}^{small}$  and  $MT_{mixed}^{medium}$  failed to show significant better scores than the Baseline translation. However, the score of  $MT_{mixed}^{large}$  is quite promising as it is significantly better than

the Baseline translation according to all the metrics. The difference between the training corpus and the test sample and the size of the training corpus are the major reasons for the lower scores of the first two models ( $MT_{mixed}^{small}$  and  $MT_{mixed}^{medium}$ ). With bigger or more similar corpora, the pre-processing model can render a better translation ( $MT_{mixed}^{large}$  and  $MT_{in-domain}$ ) than the Baseline translation.  $MT_{mixed}^{large}$  is also significantly better than  $MT_{in-domain}$ . The results reflect one important criterion in SMT training data selection: While the more the better holds here, it should also be the more similar the better. Although the in-domain corpus is much smaller than the biggest random training corpus ( $MT_{mixed}^{large}$ ), the two models trained using these two corpora work almost as well as each other (except according to TER). Therefore, we can hypothesise that if the biggest random corpus ( $MT_{mixed}^{large}$ ) was also more similar to the test corpus, the translation may get much higher scores. To sum up, the pre-processing model can improve the output of the RBMT system, especially when the pre-processing model is trained with a bigger training corpus or similar corpus.

### 7.2.3.2 Linguistic Analysis

For a human to evaluate 1000 sentences (each of which has five outputs) at both preposition level and sentence level is both time and resource costly. In addition, two models were evaluated as worse than the Baseline system. Instead of conducting a large scale human evaluation, the author first conducted a detailed examination of the translations and compared the “pseudo” English from the best pre-processing model with the original source English. To give an example of the improvements introduced by the best pre-processing model, the author compared the translation  $MT_{mixed}^{large}$  (which obtained the highest scores) with the Baseline

translation. Figure 7.4 compares the Baseline translation with the reference translation and Figure 7.5 compares the translation generated after pre-processing ( $MT_{mixed}^{large}$ ) with the same reference translation. The shaded blocks indicate where the translations are the same as the reference translation. The source English sentences are put at the top of the figures. The English sentences at the bottom are the glosses of the translations.

Source	About the processes that proactive threat scans detect							
Ref	关于	主动型	威胁	扫描	所	检测	的	进程
MT								
关于								
主动型								
威胁								
扫描								
的								
进程								
请								
检测								
Gloss	About proactive threat scans' processes please detect							

Figure 7.4: Baseline translation from the original source sentence

Source	About the processes that proactive threat scans detect							
Ref	关于	主动型	威胁	扫描	所	检测	的	进程
MT								
关于								
主动型								
威胁								
扫描								
检测								
的								
进程								
Gloss	About proactive threat scans detect DE processes							

Figure 7.5:  $MT_{mixed}^{large}$  from pre-processed source sentence

From Figure 7.4 and 7.5, we can see that although both translations share the same number of correct lexical translations (7 shaded blocks) with the reference translation, their orders are different. The gloss for the MT translation in Figure

7.4 shows that the noun phrase “proactive threat scans” is incorrectly translated as the possessor of the noun “processes”. In Figure 7.5, the attributive clause is correctly translated and is positioned before the noun “processes” it modifies. The Chinese character “DE” marks the modification relationship.

The original English sentences and the “pseudo” English sentences from the pre-processed module were compared at sentence level to reveal what changes were made by the pre-processing model to the English sentences. The following example (example 7.4) exhibits some of the changes that the pre-processing model made:

---

Example 7.4

$E_o$ : Allows other users in your network to browse files and folders on your computer.

$E_{MT}$ : Permits other user in your network to browse for the file and folder on your machine.

---

“Allows” and “computer” in the original English sentence are changed into “permits” and “machine” after pre-processing. The preposition “for” and the article “the” are two new additions found in the pre-processed English sentence. “files and folders” become singular form “file and folder”. Further qualitative assessment of these changes is necessary to reveal why, or if, these changes are leading to better translation.

Comparing  $E_o$  and  $E_{MT}$ , we observed that 998 sentences (99.8%) of all the 1000 English sentences were modified by the statistical pre-processing model, with only 2 sentences remaining unchanged. Using a function of TER, we extracted at word level the list of deletions, insertions and substitutions made to



the whole original sample by the pre-processing system. Table 7.8 reports the total number of insertions, deletions and substitutions as well as the top five most frequent changes in each category.

Category (# occurred)	Example	Frequency	
Insertion (1158)	the	248	
	will	46	
	,	41	
	”	39	
	to	36	
Deletion (992)	the	102	
	of	85	
	a	65	
	that	59	
	you	49	
Substitution (5307)	a	the	166
	can	may	150
	computer	machine	64
	that	which	58
	click	clicks	49

Table 7.8: Number and examples of insertions, deletions and substitutions

It can be seen from Table 7.8 that most of the changes are function words, for example, “*the*” is the most frequently inserted, deleted and substituted word. We are not claiming that all these changes made to the source English sentence contribute to the higher scores of pre-processed translation over the Baseline translation. However, some of the changes listed above can indeed bring improvements.

Take the insertion of punctuation marks for example, instead of using initial capitals, a term is surrounded by quotation marks in Chinese (example 7.5).

---

**Example 7.5**

$E_o$ : ...shows New Host Integrity Policy by default

**Ref:** ...默认 显示 “ 新 主机 完整性 策略 ”

**Baseline:** ...显示 新建 主机 完整性 策略 默认 情况下

**Gloss:** ...shows New Host Integrity Policy by default

$E_{MT}$ : ...displays “new Host Integrity Policy” default

$MT_{mixed}^{large}$ : ...显示 “ 新建 主机 完整性 策略 ” 默认

**Gloss:** ...shows “New Host Integrity Policy” by default

---

In example 7.5, the term in the source English ( $E_o$ ) was marked by capitalised letters while in the Chinese reference translation (Ref) a pair of quotation marks is employed to mark the term. In the pre-processed English ( $E_{MT}$ ), the term is not only capitalised but also marked by the quotation marks. Without the quotation marks in the Baseline translation, it is difficult to recognise the term, whereas the translation of the pre-processed English  $MT_{mixed}^{large}$  is more similar to the reference translation and more natural than the Baseline translation.

Deleting the pronoun “you” from an English sentence can also make its translation more similar to the reference translation as pronouns are often not translated into Chinese especially in installation documents where a series of instructions are described. In example 7.6 below, the source English sentence contains a pronoun “you” and in the Baseline it was translated into a corresponding Chinese pronoun. However, the human translation (the reference translation) does not have the translation of the pronoun. This pronoun is removed from the pre-processed English ( $E_{MT}$ ) and its translation  $MT_{mixed}^{large}$  is more similar to the reference translation.

---

Example 7.6

$E_o$ : After **you** install these client installation packages on your legacy clients...

**Ref**: 在 旧版 客户端 上 安装 这些 客户 端 安装 软件包 后 ...

**Baseline**: 在 **您** 安装 Symantec Endpoint Protection Manager.....

$E_{MT}$ : After installing these client-side installation package on the legacy client...

$MT_{mixed}^{large}$ : 在 安装 Symantec Endpoint Protection Manager...

---

In other words, the changes listed in Table 7.8 show that the pre-processing model attempted to make the English sentence more similar to the Chinese structure. The fact that  $ZH_{MT}^B$  receives higher scores than  $ZH_{MT}^F$  also reflects one of the drawbacks of most of the automatic evaluation metrics, i.e. scores of translations are based on the similarity between the machine translation output and the provided reference as the single gold standard even though other alternatives of the translation are also acceptable. Therefore, false higher scores may be generated.

The author further examined the pre-processed English sentences and divided all the “pseudo” English sentences into three groups (Table 7.9). Group one (242 sentences) contains sentences with correct English grammar and easily understandable meaning. Group two (243 sentences) consists of sentences with minor problems in English grammar and understandable meaning and group three (456 sentences) contains sentences that are ungrammatical with unclear meaning. Table 7.9 shows the distribution of the three categories.

Category	%
Correct grammar with clear meaning (example 7.7-7.8)	26
Minor error with clear meaning (example 7.9)	26
Incorrect grammar with unclear meaning (example 7.10)	48

Table 7.9: Distribution of the pre-processed sentences within three categories

The first group contains English sentences with correct grammar and with clear meaning. Most of the sentences have more or less retained the original meaning of the source English sentence (example 7.7).

---

Example 7.7

$E_o$ : You know that the **process** is safe to run in your **environment**.

$E_{MT}$ : You know that the **procedure** is safe to run in your **conditions**.

---

In example 7.7, the noun “*process*” was replaced by “*procedure*” and “*environment*” was substituted by “*conditions*”. There is no grammatical error in either sentence and the meaning has not been altered. The translations of the two English sentences ( $E_o$  and  $E_{MT}$ ) are also identical with the former showing better translation of the highlighted nouns.

However, we also found that some pre-processed English sentences in this group (20%) have different meanings from the original English sentences. For example, although both the English sentences in example 7.8 are grammatical, they do not share the same meaning with each other.

---

Example 7.8

$E_o$ : A description of the action that was taken on the risk.

$E_{MT}$ : The instruction operation which the risk adopts.

---

The translation of  $E_{MT}$  does not reflect the original meaning. This indicates that there are some degradations introduced by the pre-processing model.

The second group consists of sentences with minor grammatical errors and clear meaning, such as incorrect subject-verb agreement or missing article. In example 7.9 below, the gerund phrase in the source English sentence was substituted incorrectly by a single verb in its base form after pre-processing.

---

Example 7.9

$E_o$  : About **working with** Firewall Policies

$E_{MT}$  : About **use** firewall policies

---

Although  $E_{MT}$  is ungrammatical in example 7.9, the meaning of the pre-processed English sentence is in keeping with the original meaning and its translation is exactly the same as the reference translation. It was found that all the pre-processed English sentences in the second group retained the original meaning of the source English sentence.

The last group contains sentences that are grammatically incorrect. Moreover, the sentences are of very low comprehensibility (example 7.10). It is observed that for some of these sentences, the translation is better than the Baseline translation while for the others, the translation is worse.

---

Example 7.10

$E_o$  : Auto-Protect also reports who was logged on to the computer at delivery time.

$E_{MT}$  : The auto-protect will also report logged into to the machine on delivery time.

---

Further examination of the translations of the pre-processed English sentences in each group revealed that sentences in the second group produced more improved translations than sentences in the other two groups. Sentences in the third group generated the least number of improved translations against the baseline translation.

In summary, from the above analysis, especially Table 7.8 and the examples, we can see that the pre-processing model tries to make the source English sentence more similar to the reference translations. This resulted in some improved translations but also generated a lot of degraded translations.

Another problem of this approach is that false high automatic scores may be generated. As we explained in Chapter 2, the automatic evaluation metrics employed in this study are string-based where the word-level similarity between a candidate translation and its reference translation are measured. A translation being similar to its reference at string-level does not mean that it will be preferred by humans. Our analysis of the three groups of sentences above reveals that some sentences have been edited into totally ungrammatical and incomprehensible sentences. Their translations were not as good as the Baseline translation. A suggestion for further research is to regulate the model so that it only produces grammatical English or to regulate the RBMT system to only translate grammatical sentences but to skip the ungrammatical ones.

### **7.3 Summary**

This chapter proposes a new dictionary customisation approach and a new pre-processing method for the RBMT system, namely, an automated extracted feature-specific preposition dictionary and a statistical source reconstruction

system. Obviously, one shared advantage of these two methods is that no human intervention is needed. Both methods explore new perspectives on combining SMT and RBMT systems. The preposition dictionary makes use of the bilingual corpora while keeping the advantages of the RBMT system. The pre-processing model uses the automatic learning function of SMT to counteract the inefficiency of the RBMT system in dealing with the grammatical difference between English and Chinese. Another advantage of the proposed methods is that they are not only language independent but also system independent.

Both methods exhibit some potential benefits according to automatic evaluation metrics. However, brief examination of the translations demonstrates that there are both improvements and degradations generated in the translations. The supplementary preposition dictionary did not generate significantly better translation than the Baseline. Linguistic analysis revealed that while improvements were found in lexical translations, there were some degradations in word order. In terms of the statistical pre-processing approach, the important lesson learnt is that the bigger (or more in-domain) the corpus the better the translation of the proposed module. However, there is a risk that false high automatic scores may be reported based on the author's own analysis. How to optimise the models before applying them in production will be a major step for future work. For example, for the source reconstruction, regulating or configuring the translation process of either the RBMT system or the SMT system or both is a topic worth exploring.

## **Chapter 8: Comparison between RBMT, SMT and SPE**

How to boost the performance of the RBMT system by integrating an SMT component has been the focus of this study. In the previous chapters we have discussed that RBMT and SMT systems can be combined in many new ways and can produce better translations than the RBMT system itself. While RBMT systems are the dominant commercial systems employed by many localisation industries, recently, there is an increasing popularity of the application of SMT systems in real-life production. For example, TAUS (2009) mentioned that Autodesk decided to deploy Moses into their production cycle after their own comparison of Systran and Apertium (an open-source RBMT system) against Moses.

System comparison is of vital importance to general users and researchers. Upon introducing machine translation evaluation, we have mentioned several large scale evaluation campaigns in the MT research area (such as WMT, NIST), the purpose of which is to compare various MT systems and to advance their development. The work of Senellart et al. (2010) is perhaps the most related to the current study. They set side by side an SMT (Portage), an RBMT (Systran) and an SPE system and concluded that the SPE system was superior to both the SMT and RBMT systems. Similar to their study, this chapter is designed to compare the translation quality of the RBMT system (Systran), two SMT systems (Moses and Google) and a larger-scale SPE model with the purpose of revealing which translation model should be preferred in technical document translation, especially for achieving better translation of prepositions.



The remainder of the chapter is organised as follows. Section 8.1 introduces the experiment set-up, justifying the selection of systems and corpora. In addition, the translations from the four MT systems are obtained in this section. Section 8.2 presents the detailed preparation of the human evaluation, making use of the new selection rule proposed in Chapter 6. Evaluation results are reported along with an in-depth qualitative analysis of the strengths and weaknesses of each system in Section 8.3. Finally, Section 8.4 summarises the findings.

## **8.1 Experiment Set-up**

### **8.1.1 MT Systems**

The aim of this experiment is to find out the best system in the task of translating English IT documents into Chinese, particularly in translation of prepositions. With respect to the best MT system reported in the literature, it varies depending on the method of evaluation and size/type of training corpus. Nonetheless, Google and Systran are often evaluated as the best systems by humans (Callison-Burch et al. 2009; NIST 2008).<sup>52</sup> In fact, Google's SMT system is becoming a "standard" which is widely used in comparisons in the MT community. The advantage of Google is that it "has access to significantly greater data sources for its statistical system" (Callison-Burch et al. 2009: 10). However, unlike Moses, Google is not an open-source software which can be trained or modified by general users. What we can access is Google's pre-trained free online SMT system.

---

<sup>52</sup> NIST 2008 Open MT Evaluation – Official Evaluation Results:  
[http://www.itl.nist.gov/iad/mig/tests/mt/2008/doc/mt08\\_official\\_results\\_v0.html](http://www.itl.nist.gov/iad/mig/tests/mt/2008/doc/mt08_official_results_v0.html) [last visited 2010-04-28]

For the current comparison, the Baseline RBMT system is again Systran customised by the user dictionaries of Symantec. The Moses toolkit has been utilised to build both the pre-processing module and the post-editing module in previous chapters. However, the quality of translation from Moses as a pure SMT system has not been examined. Therefore, we decided to train a Moses SMT system using the Symantec data. We also included the translation from the online translator of Google. The difference between Moses and Google mainly lies in the training data employed. While Moses was trained on constrained data including only technical data from Symantec (i.e. homogenous domain-specific data), Google was trained on unconstrained data on various topics (i.e. heterogeneous data). Comparing Google with the tailor-made Moses can reveal how Google (with significantly larger training data) performs in English to Chinese translation compared with Moses (trained on a small amount of data). This may reveal important information on training corpus selection.

Besides the above-mentioned systems, an SPE system was also included into the comparison. The SPE system (Chapters 4 and 6) has been evaluated as being significantly better than the Baseline system Systran. However, it has not been compared with the performance of the pure Moses SMT system. We did not include the pre-processing module proposed in Chapter 7 in the comparison for the following reasons. First of all, unlike the other systems selected, the pre-processing module has not been applied in real-life production. Second, no human evaluation was conducted to cross check the results and there was a risk of false high scores based on the

author's preliminary analysis. Time and resource constraints are another important reason.

In summary, the MT systems evaluated in this study include Systran, Google, Moses and an SPE module. The detailed training and translating process of each system will be explained in the next section.

### **8.1.2 Corpora**

As mentioned before, building SMT systems requires large data sets. According to Way (2010) SMT systems are usually trained on several million words of data in order to achieve good translation quality. Therefore, the special preposition corpus set up in the chapter on methodology, which consists of 84,349 English words, seems too limited to build an effective SMT system. Although the SPE module built on this corpus (Chapter 6) demonstrated significantly better translation than Systran, we are not sure if it is suitable for the training of a pure SMT system. Additionally, in Chapter 7 we found that the pre-processing module built from the biggest corpus performed better than the one built from the preposition corpora.

Therefore, in this experiment, we continued with the use of the whole TM instead of the small preposition corpus. However, the two corpora are in the same domain in the sense that both are IT documents from the same company. The specific preposition corpus was extracted from this TM by only retaining sentences with at least one preposition.

This TM contains English and its Chinese translations aligned at sentence level. For the training of an SMT system, two corpora were generated from this TM, an English corpus and a Chinese reference corpus.

Another essential factor influencing the performance of an SMT system is the coverage of the test data in its training data (or the similarity between the training and the test data). The modifications made by an SMT system rely on the knowledge it learns from the training corpus. The more similar the training and the test set is, the better translations are expected. At the moment, no standard coverage level has been proposed for the training and test set. In our case we checked the coverage level by running the test set against the training corpus through the Analyze function of Trados (a TM tool) and reported the fuzzy matching levels calculated by Trados as a marker for the coverage level between the training and test corpus (Table 8.1).

Match levels	Sentences	Words	Percent
95% - 99%	46	794	5
85% - 94%	121	1,696	11
75% - 84%	106	1,417	9
50% - 74%	38	592	4
No Match	633	10,498	71

Table 8.1: Fuzzy matching between training and test corpora

In Table 8.1, the first column refers to how similar a test sentence and a sentence in the training corpus are. For example, 95% means that 95% of a test sentence can be matched in a training sentence. The second column refers to the number of sentences belonging to that matching level. The third and the fourth columns report the number of words and the percent of the sentences that is found in that matching level. The overall results show that for 29% of all the test sentences, more than half of each sentence (50%) can be fuzzy-matched in the training corpus. As there is no standard

as to the best fuzzy matching level at the moment, we assume that this coverage level is satisfactory.

As to the SPE system to be compared, we did not use the pre-trained SPE module discussed in Chapter 6 because it was trained on the small preposition corpus. In order to make the comparison reliable, a new SPE system was built using the bigger corpora. The English corpus was machine translated by Systran to get the raw MT output. The new SPE system was trained using the reference translations and the raw MT output. Comparing the new and old SPE modules may reveal the influence of the size of training data on the performance of SPE. The sizes of the training and tuning corpora for SMT and for SPE can be found in Table 8.2 and Table 8.3 respectively.  $Train_{EN}$  refers to the English training corpus and  $Train_{ZH}$  refers to the Chinese reference translation.  $Train_{MT}$  and  $Tune_{MT}$  refer to the Chinese translation of  $Train_{EN}$  and  $Tune_{EN}$  from Systran.

	# Sentences	# Words
$Train_{EN}$	94,622	1,250,905
$Train_{ZH}$	94,622	1,277,582
$Tune_{EN}$	944	15,884
$Tune_{ZH}$	944	13,158

Table 8.2: Training and tuning corpora for SMT

	# Sentences	# Words
$Train_{MT}$	94,622	1,212,915
$Train_{ZH}$	94,622	1,277,582
$Tune_{MT}$	944	12,496
$Tune_{ZH}$	944	13,158

Table 8.3: Training and tuning corpora for SPE

### 8.1.3 Obtaining Translations

The four systems can be divided into two groups in terms of how to obtain their translations of the test sample. One group is the “ready-to-translate” type including Systran and Google. One can upload the file and have it translated automatically by the system. The second group includes the “training-to-translate” systems, such as Moses and SPE which need a series of pre-processing steps to build up the systems first.

The training and tuning process of the Moses SMT system follows the step-by-step instructions listed on its official website (cf. Chapter 2). To build an SMT system that can translate English into Chinese, the corpora listed in Table 8.2 were used. First, the bilingual corpora were pre-processed, i.e. segmented and tokenised. Second, the Chinese language model was built using the Chinese reference corpus. Third, the system was built with the pre-processed corpora and the language model and was tuned using the tuning corpus in order to get the best output from this system. Once this was done, the system was ready to translate a new sample. The process of obtaining the translation of an SPE system for Systran has been explained in detail in Chapter 6. It follows the same steps as an SMT system but using the corpora in Table 8.3. The difference between this SPE module and the SMT system is that the SMT system can only translate English into Chinese while the SPE system can only “translate” or “post-edit” raw MT Chinese into a new version of Chinese translation.

Once the test set has been translated by the four systems, the next step is to evaluate them and compare them both at sentence level and at preposition level. The

names of the systems are used to represent their translations, i.e. Systran, Moses, Google and SPE. GTM, TER and BLEU again were employed to get the overall scores for the four systems. Human evaluation with four professional translators was also conducted to compare the four systems. The next section presents the set-up of the human evaluation.

## 8.2 Preparing Human Evaluation

Several problems were identified from results of the previous human evaluation experiment (Chapter 6). First, the inter-evaluator correlation needed to be increased; second, a great number of translations were ranked as ties (i.e. no qualitative difference between two translations). In an extreme example, one evaluator assigned 79.5% of all the translations as ties; third, the expense was considerable. The key reason for those problems was the substantial number of sentences evaluated and the number of outputs per sentence. Due to the limitation of time and funds for this evaluation, it was not feasible to have all translations evaluated by humans. Besides, how useful it is to analyse results with low inter-evaluator correlation and a large number of ties is questionable. Therefore, we made some changes to the selection process for human evaluation.

To avoid having a large number of ties, we have proposed at the end of Chapter 6 an approach to discard translations that are not easily distinguishable by evaluators. When the difference between two translations was less than 0.2 in terms of their GTM scores, there was a greater chance that humans assigned ties to the translations. Therefore, we employed GTM ( $e=1$ ) (which was found to correlate best with human evaluation at sentence level in this study) as a filter to reduce the number of

translations to be evaluated. The final selection rule used in this experiment was as follows:

Only if the difference between two GTM scores for a pair of translations differ to a value greater than 0.2 (including 0.2), will this pair of translations be included in the human evaluation.

The test in Chapter 6 showed that evaluating 11.54% of all translations using this selection rule could lead to the same conclusion with regard to the quality of the systems as evaluating all the translations. In addition, the correlation between human evaluators and between human and automatic evaluation increased slightly.

The selection rule was applied in the current experiment in the following steps with examples. No glosses are provided for the Chinese sentences since the main purpose is to show how to filter out translations through their GTM scores.

(1) Generate GTM scores at sentence level (Example 8.1)

Example 8.1		
Source	A description of the action that was taken on the risk.	GTM
Systran	在这个风险采取行动的描述。	0.470588
Moses	操作的说明针对风险所采取的。	0.888889
SPE	针对风险所采取的操作的说明。	1
Google	甲认为是在风险所采取的行动说明。	0.6

(2) Expand the four systems into pair wise comparison

There are six pairs from the four systems, namely, SPE vs. Moses, SPE vs. Systran, SPE vs. Google, Moses vs. Systran, Moses vs. Google and Systran vs. Google (Example 8.2).



Example 8.2		
Source	A description of the action that was taken on the risk.	GTM
Systran	在这个风险采取行动的描述。	0.4706
Moses	操作的说明针对风险所采取的。	0.8889
Systran	在这个风险采取行动的描述。	0.4706
SPE	针对风险所采取的操作的说明。	1
Systran	在这个风险采取行动的描述。	0.4706
Google	甲认为是在风险所采取的行动说明。	0.6
Moses	操作的说明针对风险所采取的。	0.8889
SPE	针对风险所采取的操作的说明。	1
Moses	操作的说明针对风险所采取的。	0.8889
Google	甲认为是在风险所采取的行动说明。	0.6
SPE	针对风险所采取的操作的说明。	1
Google	甲认为是在风险所采取的行动说明。	0.6

Next, compare the GTM scores of any two translations from any two systems and check if the difference between these two GTM scores meets the selection rule criteria or not. If two scores differ by more than (including) 0.2, then keep this pair for evaluation, otherwise, remove them. In example 8.2, all pairs conform to the requirement of the selection rule, except Systran vs. Google and SPE vs. Moses.

In the end, 1342 pairs (out of 3776 pairs) of translations remained in the evaluation group. A 0.2 difference value between two GTM scores for two translations means that the two translations are clearly different. Hence, the number of indistinguishable pairs could be reduced. However, we may face the challenge of losing pairs that even if they are only slightly different according to GTM, humans think are clearly different. However, Chapter 6 showed that evaluating only 11.54%

of all translations did not alter the results of evaluating all translations. Hence, it is reasonable to expect that in this comparison, evaluating 35.54% of all translations can represent the whole sample validly and reliably and that what we discarded were mostly indistinguishable pairs rather than qualitatively different pairs.

One problem that emerged is that for the same English sentences, there are cases where two translations in one pair are the same as another two translations in another pair (Example 8.3).

Example 8.3	
Source	Deleting files from the Quarantine
Systran	删除从隔离的文件
SPE	从隔离区删除文件
Google	删除从隔离的文件
Moses	从隔离区删除文件

In example 8.3, the translations from Moses and SPE are the same (Moses==SPE), and translations from Google and Systran are the same (Google==Systran). However, the GTM score differences between Systran vs. SPE and Google vs. Moses are all above 0.2. The question here concerns the final form of evaluation. In the previous evaluations, translations of the same source sentence were all put together and ranked from best to worst. Repetitive translations were usually removed. The final form looks like example 8.4 below.

Example 8.4		
Source		Deleting files from the Quarantine
Systran	Output 1	删除从隔离的文件
Google	Output 2	删除从隔离的文件
Moses	Output 3	从隔离区删除文件
SPE	Output 4	从隔离区删除文件

In the current evaluation, since we selected translations using a new method, continuing with the previous form is not appropriate. Instead, we opted for the form in Example 8.5, i.e. translations are presented in pairs along with an English sentence.

Example 8.5	
Source	Deleting files from the Quarantine
Systran	删除从隔离的文件
SPE	从隔离区删除文件

Source	Deleting files from the Quarantine
Google	删除从隔离的文件
Moses	从隔离区删除文件

To avoid judging the same translation pairs consecutively, we randomly distributed those pairs in the whole evaluation sheet. The advantage of using this approach is that the intra-evaluator correlation can be checked to see how consistent evaluators are with themselves. The only downside of this approach is that the same source sentence will be read more than once.

Besides the evaluation at sentence level, in 746 pairs of translations we also highlighted prepositions and their translations for evaluating. Pairs sharing the same translation of prepositions were not highlighted for evaluation at preposition level.

Table 8.4 below reports the number of pairs of translations for sentence level evaluation and the number of pairs for preposition level evaluation.

System ID	Preposition level	Sentence level
Systran - Google	69	147
Systran - Moses	178	299
Systran - SPE	144	221
Google - Moses	179	319
Google - SPE	141	280
Moses - SPE	35	76
Total	746	1342

Table 8.4: Number of pairs of translations for human evaluation

After randomising all pairs of translations, they were put into Excel workbooks. Due to heavy formatting and a large number of pairs, putting all the translations in one worksheet made opening the worksheet extremely slow. Therefore, it was divided into three sheets. Figure 8.1 below shows a snippet of the final evaluation sheet:

	A	B	C	D	E
1	ID	Outputs	Sentence Ranking	Preposition Ranking	Comments
2	Source	A description of the action that was taken on the risk.	Select Here		
3	Output A	针对风险所采取的操作的说明。			
4	Output B	在这个风险采取行动的 描述。			
5					
6	Source	About configuring user rights with Active Directory	Select Here		
7	Output A	关于 配置 使用 active directory 的用户 权限			
8	Output B	关于 配置 与 active directory 的用户 权利			
9					
10	Source	About infected files in the Quarantine	Select Here		
11	Output A	关于 在 隔离 的 受 感染 的 文件			
12	Output B	关于 隔离区 中 受 感染 的 文件			
13					
14	Source	About migrating unmanaged clients with CD files	Select Here	Select Here	
15	Output A	关于 有 cd 的 文件 的 迁移 的 未 处理 的 客户端			
16	Output B	关于 用 光盘 文件 迁移 非 受 管 客户端			

Figure 8.1: Sample of the human evaluation sheet

For any pair of translations, evaluators judged if “output A” is better than, worse than or equal to “output B”. As in the previous human evaluations, brief instructions written in both English and Chinese (Appendix E) together with a short questionnaire

(Appendix F) were attached in order to give general instructions and to capture data on the general working experience of the translator and their attitudes towards MT. The questionnaire can tell us whether the evaluators have mastered sufficient English and Chinese knowledge, whether they are familiar with the translated documents and whether they are biased against MT technologies or not.

### 8.3 Results

Although the number of pairs has been greatly reduced, it still took 20 hours per person to carry out the evaluation. All four evaluators were native Chinese speakers working as professional translators. From the questionnaire we learnt that they had been working as full-time translators over 6 years. The average translation throughput of each evaluator was the same, i.e. 2000 words per day. Only one evaluator had ever post-edited MT output. Two of them mentioned that MT could reduce the effort of translation. In the *Comments* column of the evaluation sheet, all four evaluators made some comments, either explaining the meaning of his/her notation such as “when ‘equal’ is selected, that means both of them are not readable”, or suggesting a correct translation, for example one evaluator noted down that “custom should be translated into ‘自定义’”.

This section explores and analyses the results in order to address the research aim of this experiment, i.e. to reveal the advantages and disadvantages of the four systems in English to Chinese translation, especially in translation of English prepositions. To begin with, the inter-evaluator correlation and the intra-evaluator correlation are assessed in Section 8.3.1.

### 8.3.1 Inter- and Intra-evaluator Correlation

One expectation of the selection rule was that it could ensure higher inter- and intra-evaluation correlation indicating more reliable results. For the 1342 pairs of translations which were selected for sentence level evaluation, 1004 pairs received the same judgement by at least three evaluators. A breakdown of pair-wise correlation between evaluators can be found in Table 8.5. H1 to H4 below represents the four human evaluators respectively.

	AGREEMENT LEVEL	KAPPA
H1-H2	62.14%	0.4398
H1-H3	61.99%	0.4428
H1-H4	76.55%	0.6465
H2-H3	65.07%	0.4594
H2-H4	65.07%	0.4794
H3-H4	65.97%	0.4947
Average	66.13%	0.4905

Table 8.5: Inter-evaluator correlation

Overall, the Kappa correlation of the four evaluators is 0.4905 (moderate correlation). It has been pointed out earlier that in general a commonly reported inter-evaluator correlation score for ranking evaluation falls into the interval [0.21-0.40] indicating a Fair agreement. For example, the inter-evaluator correlation of sentence-level ranking reported by Callison-Burch et al. for the WMT evaluation campaigns held in 2007, 2008 and 2009 were 0.373, 0.367, 0.323 respectively, all of which are fair agreements. The inter-evaluator agreement level for our pilot project was only 0.273 (cf. Chapter 6). Therefore, the inter-evaluator correlation in this experiment is noticeably higher than the ones mentioned above. In addition, unlike the results in Chapter 6 where one evaluator had extremely low correlation with the others, all the evaluators in this experiment have more or less the same level of

agreement with each other. In other words, using the selection rule can enhance the correlation between human evaluators indicating more valid results.

At preposition level, all together 746 pairs have prepositions highlighted, 543 pairs received unanimous judgement by at least three evaluators. The Kappa value of the inter-evaluator agreement at preposition level is 0.4285 which is also a moderate correlation value.

The intra-evaluator correlation of each evaluator was calculated using the 109 pairs of translations each of which was randomly repeated in the whole sample. Table 8.6 below presents the intra-evaluator correlation of each evaluator. The four evaluators have on average a substantial intra-evaluator agreement, with one evaluator (H3 in Table 8.6) having an almost perfect agreement with himself/herself.

	AGREEMENT LEVEL	KAPPA
H1	88.99%	0.7798
H2	85.32%	0.7064
H3	90.83%	0.8165
H4	88.07%	0.7615

Table 8.6: Intra-evaluator correlation

It could be concluded that by presenting the evaluation in the form of pair-wise comparison, and especially by discarding pairs of translations using our 0.2 rule, not only did we save time and cost but we also made the human evaluation task easier and obtained more valid and consistent results.

## 8.3.2 System Level Comparison

### 8.3.2.1 Results of Automatic Evaluation

The translation of the systems was first measured and compared by the automatic evaluation metrics, namely, GTM (e=1), GTM (e=1.2), BLEU and TER. Readers may remember that the GTM (e=1) and GTM (e=1.2) differ from each other in their penalty to the word order difference between a translation and a reference translation. GTM and BLEU score ranges from 0 to 1, indicating how similar a candidate translation is to the reference translation with 1 representing a perfect match. The higher a GTM or BLEU score, the better a translation. On the other hand, the higher a TER score, the poorer a translation, with a score of 0 representing no edit is needed for a translation compared to its reference. Table 8.7 reports the automatic scores of the sample translation from the four systems. According to all the metrics, the overall ranking of the four systems is consistent, i.e. SPE is better than Moses both of which are better than Systran and Google.

SYSTEM ID	GTM (e=1)	GTM (e=1.2)	BLEU	TER
Google	0.388	0.330	0.212	0.557
Systran	0.417	0.357	0.241	0.538
Moses	0.559	0.483	0.438	0.405
SPE	<b>0.561</b>	<b>0.487</b>	<b>0.448</b>	<b>0.399</b>

Table 8.7: Automatic evaluation scores for the four systems

We performed significance tests on the scores of any two systems using statistical re-sampling (Koehn, 2004). Test results show that Moses and SPE are significantly better than the Baseline translation and than Google translation (at  $p < 0.01$ ). Google is not as good as the Baseline translation; on the contrary, the Baseline is significantly



better than Google translation ( $p<0.05$ ). Although the score of SPE is higher than that of Moses, the two results do not demonstrate statistically significant difference.

### 8.3.2.2 Results of Human Evaluation

Human evaluation results were then summarised to cross check the results of automatic evaluation. Since the human evaluation was conducted in a pair wise form, for any pair of systems, we summarised the percent of times that one system was judged to be better than the other system in a pair by at least three evaluators. Table 8.8 reports how often the column system was evaluated as **better than** the row system by the majority of human evaluators.

	Google	Moses	SPE	Systran
Google	/	58.57%	<b>71.69%</b>	41.07%
Moses	7.14%	/	<b>40%</b>	14.1%
SPE	4.57%	31.67%	/	4.55%
Systran	30.36%	54.62%	<b>80.11%</b>	/

Table 8.8: Percent of times that the column system is judged as better than the row system (sentence level)

From Table 8.8 we can see that SPE is judged as better than any other system (Google, Systran and Moses) most frequently, followed by Moses. Systran is better than Google. The ranking of the four systems inferred here from best to worst is: SPE>Moses>Systran>Google. “>” means the left system is better than the right system according to human evaluation. However, statistical significance tests show that the difference between SPE and Moses is not significant; hence, no conclusive distinction can be made between the quality of SPE and Moses. While the rest of any two systems is significantly different at  $p<0.01$ , the difference between Systran and Google is less significant at a low cut-off point ( $p<0.05$ ).

### 8.3.2.3 Correlation between Human and Automatic Metrics

With regards to the overall ranking of the four systems, results of human and automatic evaluation all concurred that the best to worst system ordering was SPE>Moses>Systran>Google. In addition, statistical tests on both human evaluation results and automatic evaluation scores showed that there was no significant difference between SPE and Moses.

To check the correlation at sentence level is more complicated than at the system level. Since the pairs were selected according to GTM ( $e=1$ ) scores, to avoid bias against other automatic metrics, only the correlation between human evaluation and GTM scores was examined. GTM assigns scores to translations in a pair, while humans selected the better translation in the pair. In Chapter 6 (section 6.2.3.3), we suggested transforming the automatic scores into the same form of results as that of the human evaluators. By regarding GTM as a special “evaluator”, we can calculate both the agreement level and Kappa score. The agreement level between GTM and a human evaluator was obtained by summarising the total number of pairs where automatic evaluation reached the same decision as the human evaluation (i.e. the higher score translation in a pair was also judged as better by the human evaluators), then divided by the total number of pairs compared. The Kappa score was calculated through the statistical template (Chapter 5). Table 8.9 reports the results.

	AGREEMENT LEVEL	KAPPA
H1-GTM	49.07%	0.2728
H2-GTM	58.43%	0.3315
H3-GTM	63.58%	0.3832
H4-GTM	51.64%	0.2903
Overall	55.68%	0.3195

Table 8.9: Correlation between human evaluation and GTM

In general, automatic evaluation metrics were consistent with human evaluation (Fair agreement). The agreement levels tell us that humans and automatic metrics agreed with each other on more than 50% of the pairs. However, there are still many translations where the judgement of humans and automatic metrics are different. The main reason for the discrepancy between human and automatic evaluation is that all translation pairs differ according to GTM but humans still evaluated some of them as ties. Nonetheless, the ratios of ties compared to the results in Chapter 6 are much smaller due to the filtering rule (on average 29% of translations were ties). Callison-Burch et al. (2007, 2008 and 2009) pointed out in their analysis that automatic metrics generally assign real numbers as scores while humans may easily assign ties which might create a bias against the correlation between automatic metrics and human evaluation. Therefore, they usually excluded pairs ranked as ties by human evaluators. In the current study if we exclude pairs that human evaluators ranked as ties, the agreement levels and correlations between human and automatic evaluation increased greatly to a moderate correlation (Table 8.10).

	Agreement level	Kappa
H1-GTM	81.82%	0.6358
H2-GTM	77.30%	0.5461
H3-GTM	77.67%	0.5532
H4-GTM	80.94%	0.6182
Average	79.43%	0.5883

Table 8.10: Refined correlation between human and GTM

Compared to the correlation reported in Section 6.2.3.3 where removing the ties that humans assigned generated just a fair correlation between humans and GTM and

the agreement level was only 66%, we can conclude that this filtering rule can boost the correlation between human and automatic evaluation.

Overall, humans and automatic metrics agreed with each other as to the ranking of the four systems and the judgements at sentence level. The filtering rule not only increases the inter-evaluator agreement but also increases the agreement between automatic and human evaluation.

### 8.3.3 Preposition Level Comparison

Following the methods employed in Section 8.3.2.2, we calculated the percent of times that one system was ranked as better than the other systems by at least three evaluators at preposition level. Table 8.11 reports the percent of times the column system was judged as better than the row system by the majority of human evaluators.

	Google	Moses	SPE	Systran
Google	/	73.33%	<b>78.07%</b>	43.75%
Moses	15.83%	/	<b>45.45%</b>	14.75%
SPE	15.79%	40.91%	/	7.7%
Systran	35.42%	72.13%	<b>82.91%</b>	/

Table 8.11: Percent of times that the column system was judged as better than the row system (preposition level)

The ranking of the four systems from best to worst is consistent with their ranking summarised in Section 8.3.2.2, i.e. SPE>Moses>Systran>Google. Similarly, statistical significance tests show that there is no significant difference between SPE and Moses; Moses and SPE are significantly better than Systran and Google. The difference between Systran and Google is not significant either in terms of their translation of prepositions.

### 8.3.4 Linguistic Analysis

In Chapter 7 (see Section 7.1.2), two basic error types (i.e. lexical translation and word order) reported by the internal users of Symantec were employed to compare the translations of various systems. We employed these two categories again to briefly compare the four systems. Only examples where an agreement was reached by the majority of the evaluators were studied. Moreover, insignificantly different pairs (e.g. Moses vs. SPE or/and Google vs. Systran) were not considered.

#### 8.3.4.1 Sentence Level Comparison

At sentence level, based on the results of both human and automatic evaluation, significantly different pairs include: Systran>Google, Moses>Systran, SPE > Systran, Moses > Google, SPE > Google. This section explores the question of in which of the two categories defined above one system was evaluated as better than the other in a pair. For example, in which aspect does Systran perform better than Google? To answer the question, for each of the five pairs, the author calculated the number of instances of the two categories (Table 8.12). Before we concentrate on the numbers in Table 8.12, let us first look at several examples of the two categories.

	Lexical Translation	Word Order
Systran > Google	26	28
Moses > Systran	109	51
SPE > Systran	130	36
Moses > Google	118	54
SPE > Google	134	94

Table 8.12: Frequency of the two categories at sentence level

The first category is lexical translation (example 8.6). In all the examples below, differences between the two translations are highlighted.

---

#### Example 8.6

**Source:** On the Monitors tab, click the Logstab.

**Systran:** 在 监视器 上 请 选中 , 单击 logstab 。 /pīnyīn: xuǎn zhòng/

**Gloss:** On Monitors please select, click Logstab.

**Moses:** 在 “ 监视器 ” 选项卡 上 , 单击 “ logstab 。 /pīnyīn: xuǎn xiàng kǎ/

**Gloss:** On “Monitors” tab on, click Logstab.

---

Moses was evaluated as better than Systran in this sentence by the majority of evaluators. The first difference between the two translations is the addition of the punctuation marks in the Moses output. As we mentioned in Chapter 7 in Chinese a term is usually marked by quotation marks. The most important difference between the two translations lies in the translation of the word “*tab*”. Systran analysed it as a verb and translated it into a verb in Chinese (which can be back translated into “please select or check”). Moses correctly translated the word into a noun (meaning “an option”). Readers may have noticed that while there are some improvements in the translation of Moses, there is also degradation. For example, the single quotation mark added at the end of the translation in Moses is incomplete and hence, incorrect. But overall, in this example, Moses was evaluated as better than Systran mainly because of better lexical selection.

The second category is word order including any form of order changes (Example 8.7).

---

Example 8.7

**Source:** Check each computer on **which you want to install client software**.

**SPE:** 选中 **您要安装客户端软件的** 每台计算机上。

**Gloss:** Check you want install client software DE (modifier marker) each computer on.

**Google:** 检查 每个 计算机 上 **要安装客户端软件**。

**Gloss:** Check each computer on want install client software.

---

In the source English sentence, there is an attributive clause. It has been mentioned that when translating an attributive clause, it should occur before the noun it modifies in Chinese. From the position of the highlighted words we can see that SPE conforms to this rule. This is perhaps the reason why the majority of human evaluators preferred the translation of SPE than that of Google for this example.

From Table 8.12 we can see that first, compared to Google which is trained using unrestrained large size of data, the customised version of Systran with its own user dictionaries performs better both in terms of lexical selection and word order. This exemplifies to some extent the importance of the customised user dictionary.

Second, SPE and Moses are better than Systran mainly at lexical translation. The finding can help us understand why automatic metrics favour SMT output over RBMT output while humans prefer RBMT output to SMT output. This is because SMT and SPE tend to have more similar lexical translations to the reference than the Systran output does. Since most automatic metrics (especially string-based ones) depend on n-gram matching between a translation and its reference, high scores may be generated. However, humans may pay more attention to the structures and favour the RBMT translations which are generated based on linguistic rules.

Thirdly, SPE generated more correct lexical translations than Moses while Moses has more correct word order than SPE when both are compared to the Baseline Systran. Finally, compared to Google, SPE and Moses are better than Google mainly at lexical translation as well.

#### 8.3.4.2 Preposition Level Comparison

At preposition level, only four pairs are significantly different from each other and they are Moses>Systran, SPE>Systran, Moses>Google and SPE>Google. A comparison of the translations of prepositions from any two systems is reported in Table 8.13. Similarly to the above sentence level analysis, we want to explore in which aspect one system was evaluated as better than the other.

	Lexical Translation	Word Order
Moses>Systran	65	26
SPE>Systran	80	22
Moses>Google	60	37
SPE>Google	52	45

Table 8.13: Frequency of the two categories at preposition level

For preposition translations, the biggest difference between any two systems is also lexical translations. In terms of lexical translation of prepositions, several sub-categories can be identified based on the error typology we set up at the beginning of the study. For example, in some pairs, it is the translation of single prepositions that is different (Example 8.8).

---

#### Example 8.8

**Source:** adding **to** a policy

**Systran:** 添加 **对** 策略 /pīnyīn: duì/

**SPE:** 添加 **到** 策略 /pīnyīn: dào/

---



Both translations can be glossed back into the same English preposition *to*. However, the two Chinese translations are different in meaning. The first one usually occurs in contexts such as “*To me, this question is too hard to answer.*” The second translation is correct in the example by meaning “a destination of an action”. The majority of evaluators all agreed that SPE is better than Systran for this sentence.

In some other pairs, the two translations differ in their translation of English preposition into Chinese circumposition (Example 8.9). While Google failed to output the second part of the circumposition, Moses not only produced a correct translation of the preposition but also correctly added punctuation around the term. Moses was judged as better than Google by the evaluators in this sentence.

---

Example 8.9

**Source:** In Description, type a description of the new Application and Device Control Policy.

**Google:** 在说明，键入说明新的应用和设备控制政策。/pīnyīn: zài/

**Gloss:** In Description, type description new Application and Device Control Policy.

**Moses:** 在“说明”中，输入新应用程序与设备控制策略的说明。/pīnyīn: zài...zhōng /

**Gloss:** At “Description” at, type new Application and Device Control Policy DE (genitive) description.

---

Fixed phrases are prevalent in this corpus. Sometimes, different translation systems generate different translations (example 8.10).

---

Example 8.10

**Source:** **Based on** the existing security policies, you may or may not have more than one location available.

**Google:** 凭 现有的 安全性 策略 , 您 可以 或 不 可以 有 超过 可 用 一个 的 位置 。 /pīnyīn: píng/

**Gloss:** Based on...

**Moses:** 根据 现有的 安全 策略 , 您 有 可能 不 能 提供 一个 以上 的 位置 。 /pīnyīn: gēn jù/

**Gloss:** Based on...

---

Although both translations correspond to the English phrase “*based on*”, the translation of Google (one Chinese character) is not as accurate and natural as the translation of Moses (two Chinese characters) in this context and the majority of evaluators judged Moses as better.

As discussed in Chapter 3, PP attachment is a big challenge to any MT system. The error associated with this structure is mainly incorrect word order. From the comparison, we can see that SPE module can generate better translation with correct word order than Systran does (example 8.11).

---

Example 8.11

**Source:** Repeat this procedure for all reporting servers.

**SPE:** 对 所有 报告 服务器 重复 此 过程 。

**Gloss:** For all reporting servers repeat this procedure

**Systran:** 重复 全部 报告 服务器 的 这个 过程 。

**Gloss:** Repeat all reporting servers DE (genitive) this procedure.

---

The translation from the SPE module attached the preposition phrase to the verb; however, Systran analysed the preposition phrase as a modifier of the noun. Different analysis results in different word order of the prepositional phrase. In this example, the output of SPE is correct and its translation was evaluated better by the majority of human evaluators.

## **8.4 Summary**

This chapter compares the performance of Systran, Moses, SPE and Google. Overall, no significant difference was reported between Moses and SPE which were trained on the same size of corpora, although both the automatic metrics and human evaluations marginally favoured SPE. The reason for the superiority of the SPE system is that it makes use of both the advantages of the RBMT and SMT systems.

Moses and SPE are significantly better than Google. Remember that Google in this study was not trained using the same corpora as Moses and SPE. We mentioned at the beginning of the chapter that in large-scale evaluation campaigns where Google was trained on the same data as the other participating systems, Google was found to be the best system on many occasions. The finding confirms the importance of in-domain training data on the performance of an SMT system.

Systran is better than Google at sentence translation but not significantly better at preposition translation. One determining factor for this is the customised in-domain dictionaries of Systran.

Another important finding of the results is that the selection rule for human evaluation could not only save time and cost but also boost the inter-evaluator correlation, and the correlation between GTM and human evaluation.

## Chapter 9: Conclusion

It was established from existing literature and the error report of internal translators of Symantec that prepositions were a major challenge faced by the RBMT system when translating to Chinese due to the fact that English prepositions are polysemous and the translation correspondents between the two languages are variable. The aims of the current study were to identify the errors of the MT translation of prepositions in IT-domain documents from English into Chinese, to explore approaches to improve and to outline the best systems in terms of translation of prepositions. While previous research so far has tended to focus on translation into English rather than from English into Chinese, the focus of this study was MT of English prepositions into Chinese. In this respect, we hope to have made a contribution to the knowledge regarding the machine translation of prepositions from English into Chinese. This thesis sought to address the following research questions (RQ).

(RQ1) Which prepositions in the Symantec corpus are translated unsatisfactorily?

(RQ2) Which errors occur most frequently in our selected corpus?

(RQ3) What types of errors are associated with each preposition?

(RQ4) What existing solutions are suitable for tackling the most common errors?

(RQ5) What are the possible effective solutions that have not yet been tested?

Before examining which prepositions were handled unsatisfactorily by the RBMT system of the study, in Chapter 4 several exploratory pilot tests on some existing approaches such as Controlled Language (CL) authoring, User Dictionary (UD), statistical post-editing (SPE) and automatic Search & Replace (S&R) were tested (RQ4). Chapter 5 then answered the first three questions through a detailed human

evaluation of the translation of prepositions (RQ1, RQ2 and RQ3). As an answer to RQ5, three new approaches were proposed to improve the translation of prepositions.

## 9.1 Important Findings

The results of the first human evaluation (Chapter 5) revealed that the RBMT system failed to generate satisfactory translations for around 50% of the prepositions. However, the other half of the machine generated translations for prepositions did not need further post-editing. In addition, some prepositions seemed to be handled better (such as *about* and *as*) than others (such as *in* and *with*). We therefore identified the most problematic English prepositions for Systran in this context.

Five basic error types were identified from the translation of prepositions. They were Incorrect Position-Word Level, Incorrect Position-Phrase Level, Incorrect Lexical Selection, Incomplete Translation and Translation Missing. Further examination of the errors in the translation of each preposition revealed that the most common types of errors vary across prepositions. Incorrect Lexical Selection was the most frequent error in the translation of prepositions *as*, *with*, *for* and *to*. Incomplete Translation was most often found in the translation of prepositions *in* and *on*. Position error was the biggest problem for prepositions *from* and *by* while Translation Missing was prevalent in the translation of the preposition *of*. This finding reveals that different types of error are associated with different prepositions.

The first attempt to reduce the errors was to explore and modify the process of Statistical Post-Editing (SPE) (Chapter 6). Constraining the SPE to be preposition-specific by only retaining phrases containing prepositions failed to outperform the unmodified SPE module. If the phrase table was constrained too much,

the SPE module would not be able to learn as much information as the unmodified module to post-edit the raw RBMT output. On the other hand, if the phrase table was loosely constrained, then it was almost the same as the unmodified SPE and might not be preposition-specific enough. An important lesson learnt from this experiment was that the translation of prepositions was not isolated but closely depended on related information from the other sentence constituents. That is why we proposed a general pre-processing approach in Chapter 7.

Both human evaluation and automatic evaluation concurred that the unmodified general SPE module was significantly better than the Baseline RBMT system both at sentence translation and preposition translation. Comparing the translation from the unmodified SPE and the Baseline system showed that the most frequently corrected preposition error by an SPE module was Incorrect Position. Incomplete translations of prepositions, especially *in* or *on* were the second most frequently corrected error. However, there were also degradations generated as 21% of the prepositions were translated less well by the SPE system than the baseline RBMT system.

We also proposed building an automated preposition dictionary and a statistical source pre-processing method in Chapter 7. The first was to obtain a supplementary preposition dictionary for the RBMT system automatically extracted from the phrase table of an SMT system. The second was to edit the source texts automatically to better suit the RBMT system. All automatic evaluation metrics reported a slight improvement of the translation with the supplementary preposition dictionary and a significant 10% increase in scores after source pre-processing. This finding demonstrates that a preposition-specific dictionary does not significantly improve the

translation of prepositions into Chinese, but that source pre-processing is worthy of future research as a way of eliminating errors.

The last chapter compared Systran, the Moses system, Google translator and an SPE module for Systran to examine which translation architecture should be preferred, especially for achieving better translation of prepositions. This led us to conclude that the best systems were the Moses SMT system and the Systran + SPE module, which are not significantly different from each other. However, both the automatic and human evaluation results slightly favoured the SPE module which took advantage of the RBMT and the SMT system. At the moment, both paradigms have been used in real life localisation contexts, for example, Symantec has employed the SPE module in their production and Autodesk uses Moses in theirs.

Systran was evaluated as better than Google where the superiority of Systran was mainly demonstrated at the level of lexical translation, with a few improved cases in word order. This confirms the importance of the domain-specific user dictionaries that Symantec compiled. However, as pointed out earlier, we should not rush to a conclusion that Google is not as good as Systran because the Google translator compared in this study was not trained on the Symantec corpus. In other public evaluation campaigns where Google was trained using the in-domain corpus, it was also often evaluated as the best system. For example, in the NIST 2008 official evaluation campaign, Google's SMT system was evaluated as the best system in English to Chinese translation.

Extensive evaluation has been conducted during this research, which enhanced our understanding of human evaluation and the correlation between automatic



evaluation and human evaluation. The most important contribution of the current study is that it opens up a new research direction with regard to the relation between automatic evaluation and human evaluation. Instead of arguing which evaluation method is better or which automatic metric correlates best with human evaluation, the current study suggests using automatic metrics to help increase the reliability of human evaluation. A filtering rule was put forward and the application of this approach in our later experiments proved that this approach would be both a time and resource saver for the research community and to commercial users.

Although most automatic metrics were designed to measure corpus-level quality, sentence level correlation (especially language-specific correlation) has attracted much attention recently. One conclusion reached based on the results of the study is that in terms of Chinese IT document evaluation at sentence level, GTM ( $e=1$ ) was found to correlate better with human evaluation than BLEU and TER.

Various ways of training SMT systems were involved in this study. In the pre-processing experiments, we found that although the best system was trained using the biggest mix-domain (heterogeneous) corpus, the system trained using the much smaller in-domain (homogeneous) corpus was almost as good as the best system. Hence, the importance of the similarity between the training and the test sample is established. This is shown more clearly by the fact that Google failed to outperform Moses. In other words, the bigger the in-domain corpus, the better, and the more similarity between the training and test sample, the better.

One shared advantage of the new approaches is that no human intervention and no tailor-made rules were necessary. On the other hand, automatically modifying

everything did not always result in improvements and the performance of these approaches was influenced by the coverage or domain of the training corpus. Detailed qualitative and quantitative analysis of the translations revealed that there were both benefits and drawbacks to all of the approaches. Nonetheless, these findings enhance our understanding of the translation of prepositions in IT-domain documents. Moreover, multiple new perspectives on combining the state-of-the-art MT systems were presented.

The findings of the study benefits Symantec in terms of revealing the in-depth problems of preposition translations, suggesting the most suitable automatic evaluation metric for Chinese evaluation, how to conduct human evaluation more effectively and what types of corpora should be used in building an SPE module.

## 9.2 Limitations and Future Research

This study is limited in several ways. To start with, the generalisability of the findings to other prepositions or other contexts might be questioned due to the fact that only the top **ten** most frequently occurring prepositions in the IT-domain documents were studied. However, the top ten prepositions account for 90% of all the preposition instances. In addition, the distribution of these ten prepositions in our corpus is more or less the same as that in other corpora such as the Penn Treebank, COBUILD and LOB where nine out of the ten prepositions are the same with the only exception being the preposition *about* which is particularly frequent in this corpus). Therefore, we can assume that our findings would have some general applicability to other domains or documents.

Another limitation is that the focus of the study is on the performance of Systran

which is a proprietary system customised by Symantec. The customisation level might affect the final findings of the study. Although we argued that there were few entries in the UD of Symantec related to prepositions and hence, the UD does not influence the study greatly, since various companies have their own customisation level, the generalisability of the findings to other types of documents, to other RBMT systems and to other localisation companies is definitely a topic worthy of further study.

Another problem is that only three of the commonly used string-based evaluation metrics were employed and examined in this study. As discussed in the literature review, other more complex metrics also exist. To continue the research we have initiated here about the proposed new function of automatic metrics is also of great practical importance. In addition, results and feedback from our human evaluation suggested that pair wise ranking (compared to ranking of multiple translations) may be one effective way to improve the reliability of human evaluation.

Some other interesting future directions have already been pointed out at the end of each chapter. For example, there were some unique improvements that were only found in one of the modified SPE modules. To integrate the improvements of the two modules together may ensure much better translation. Both the automated dictionary customisation approach and the source pre-processing module could also be optimised in order to obtain better translation. Finally, another advantage of the approaches proposed in this study is that they are system- and language-independent and they are compatible with each other. Hence, combining all the pre- and post-processing approaches may be another important direction of research.

### **9.3 Closing Words**

Taken together, the findings add to the current literature on the translation of prepositions from English into Chinese of IT documents in a localisation context, on the correlation between automatic and human evaluation and on system combination. The implication of the study is that no single approach can tackle all the problems of MT systems; instead, a more practical approach is to pursue incremental improvements. With regards to this academic and industry collaboration, the evaluation experiments were usually constrained by the budget, resources and the usefulness to practical implementation. However, overall, the researcher benefited greatly from this collaboration being able to access both academic expertise and cutting-edge technologies and rich resources in real-life contexts.

## References

- Agarwal, A. and Lavie, A. 2008. Meteor, M-BLEU and M-TER: evaluation metrics for high-correlation with human rankings of machine translation output. *IN: Proceedings of the Third Workshop on Statistical Machine Translation*, 19 June, Columbus, Ohio, pp.115-118.
- Albrecht, J. and Hwa, R. 2007. A re-examination of machine learning approaches for sentence-level MT evaluation. *IN: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, 23-24 June, Prague, Czech Republic, pp.880-887.
- Alegria, I., Casillas, A., Ilaraza, A.D., Igartua, J., Labaka, G., Lersundi, M., Mayor, A., Sarasola, K., Saralegi, X. and Laskurain, B. 2008. Mixing approaches to MT for Basque: selecting the best output from RBMT, EBMT and SMT. *IN: Proceedings of Mixing Approaches to Machine Translation (MAMT 2008)*, 14 February, Donostia-San Sebastian, Spain, pp.27-34.
- Allen, J. and Hogan, C. 2000. Toward the development of a post-editing module for raw machine translation output: a controlled language perspective. *IN: Proceedings of the 3rd International Workshop on Controlled Language Applications (CLAW 2000)*, 29-30 April, Seattle, Washington, pp.62-71.
- Antonopoulou, K. 1998. Resolving ambiguities in Systran. *Translation Service*. 98. [Online] Available from: <http://api.ning.com/files/> [Accessed 16 June 2010]
- Aranberri, M.N. 2009. *-ing Words in RBMT: Multilingual Evaluation and Exploration of Pre- and Post-processing Solutions*. PhD thesis. Dublin City University.
- Arnold, D. 2003. Why translation is difficult for computers. *IN: Somers, H. (ed.) Computers and translation: A translator's guide*. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp.119-142
- Arnold, D.J., Balkan, L., Meijer, S., Humphreys, R.L. and Sadler, L. 1994. *Machine Translation: An Introductory Guide*. London: Blackwells-NCC.
- Attnäs, M., Senellart, P. and Senellart, J. 2005. Integration of Systran MT systems in an open workflow. *IN: Proceedings of the 10th Machine Translation Summit*, 12-16 September, Phuket, Thailand, pp.211-218.

- Babych, B., Hartley, A. and Sharoff, S. 2009. Evaluation-guided pre-editing of source text: improving MT-tractability of light verb constructions. *IN: Proceedings of the 13th Annual Conference of the European Association for Machine Translation (EAMT 2009)*, 14-15 May, Barcelona, pp.36-43.
- Banerjee, S. and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *IN: Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization of the Annual Meeting of Association of Computational Linguistics (ACL 2005)*, 29 June, Ann Arbor, Michigan, pp.65-72.
- Biber, D. 1990. Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing*. 5(4), pp.257-269.
- Biber, D. 1993. Representativeness in corpus design. *Literary and Linguistic Computing*. 8(4), pp.243-257.
- Boslaugh, S. and Watters, P.A. 2008. *Statistics in a Nutshell*. O'Reilly Media, Inc., the United States of America.
- Bowker, L. and Pearson, J. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora*. London; New York, Routledge.
- Brill, E. and Resnik, P. 1994. A rule-based approach to prepositional phrase attachment disambiguation. *IN: Proceedings of the 15th International Conference on Computational Linguistics (COLING94)*, 5-9 August, Kyoto, Japan, pp.1198-1204.
- Brown, P.F., Cocke, J., Della Pietra, S.A., Della Pietra V.J., Jelinek, F., Lafferty J.D., Mercer R.L. and Roossin, P.S. 1990. A statistical approach to machine translation. *Computational Linguistics*. 16(2), pp.79-85.
- Brown, P.F., Della Pietra, V.J., Della Pietra, S.A. and Mercer R.L. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*. 19(2), pp.263-312.
- Cahill, A. 2009. Correlating human and automatic evaluation of a German surface realiser. *IN: Proceedings of the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing of the 47th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP 2009)*, 2-7 August, Singapore, pp.97-100.
- Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. and Schroeder, J. 2007. (Meta-)evaluation of machine translation. *IN: Proceedings of the 2nd Workshop*

- on *Statistical Machine Translation*, 23-24 June, Prague, Czech Republic, pp.136-158.
- Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. and Schroeder, J. 2008. Further meta-evaluation of machine translation. *IN: Proceedings of the 3rd Workshop on Statistical Machine Translation (WSMT 2008)*, 19 June, Columbus, Ohio, pp.70-106.
- Callison-Burch, C., Koehn, P., Monz, C. and Schroeder, J. 2009. Findings of the 2009 workshop on statistical machine translation. *IN: Proceedings of the 4th Workshop on Statistical Machine Translation (WMT 2009)*, 30-31 March, Athens, Greece, pp.1-28.
- Callison-Burch, C., Osborne, M. and Koehn, P. 2006. Re-evaluating the role of BLEU in machine translation research. *IN: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, 3-7 April, Trento, Italy, pp.249-256.
- Carl, M. and Way, A. (eds.) 2003. *Recent Advances in Example-Based Machine Translation*. Dordrecht/ Boston / London: Kluwer Academic Publishers.
- Carletta, J. 1996. Assessing agreement on classification tasks: the Kappa statistic. *Computational Linguistics*. 22(2), pp.249-254.
- Carpenter, W.T., James, I.K., Bilbe, G. and Bischoff, S. 2004. At issue: a model for academic/industry collaboration. *Schizophrenia Bulletin*. 30(4), pp.997-1004.
- Chen, K.H. and Chen, H.H. 1996. A hybrid approach to machine translation system design. *Computational Linguistics and Chinese Language Processing*. 1(1), pp.159-182.
- Chen, Y., Eisele, A., Federmann, C., Hasler, E., Jellinghaus, M. and Theison, S. 2007. Multi-engine machine translation with an open-source decoder for statistical machine translation. *IN: Proceedings of the 2nd Workshop on Statistical Machine Translation*, 23-24 June, Prague, Czech Republic, pp.193-196.
- Cohen, L., Manion L. and Morrison K. 2007. *Research Methods in Education*. Taylor and Francis Ltd. Sage Publication Ltd.
- Costa-Jussà, R.M., Farrús, M., Mariño, B.J and Fonollosa, A.R.J. 2010. Automatic and human evaluation study of a rule-based and a statistical Catalan-Spanish machine translation systems. *IN: Proceedings of the International Conference on Language Resources and Evaluation*, 17-23 May, Valletta, Malta, pp.1706-1711.

- Coughlin, D. 2003. Correlating automated and human assessments of machine translation quality. *IN: Proceedings of MT Summit IX*, 23-27 September, New Orleans, Louisiana, pp.63-70.
- Crego, J.M. and Marino, J.B. 2006. Integration of POSTag-based source reordering into SMT decoding by an extended search graph. *IN: Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2006)*, 8-12 August, Boston Marriott, Cambridge, Massachusetts, pp.29-36.
- Dabbadie, M., Hartley, A., King, M., Miller, K.J, Hadi, W.M., Popescu-Bellis, A., Reeder, F. and Vanni, M. 2002. A hands-on study of the reliability and coherence of evaluation metrics. *IN: Proceedings of Workshop on Machine Translation Evaluation: Human Evaluators Meet Automated Metrics at the Third International Conference on Language Resources and Evaluation (LREC 2002)*, 27 may, Las Palmas, Canary Islands, Spain, pp.8-16.
- Doherty, S. and O'Brien, S. 2009. Can MT output be evaluated through Eye Tracking? *IN: Proceedings of Machine Translation Summit XII*, 27-29 August, Ottawa, Canada, pp.214-221.
- Dugast, L., Senellart, J. and Koehn, P. 2007. Statistical post-editing on SYSTRAN's rule-based translation system. *IN: Proceedings of the 2nd Workshop on Statistical Machine Translation (WSMT 2007)*, 23 June, Prague, Czech Republic, pp.220-223.
- Dugast, L., Senellart, J. and Koehn, P. 2009. Statistical post editing and dictionary extraction: Systran/Edinburgh submissions for ACL-WMT2009. *IN: Proceedings of the 4th Workshop on Statistical Machine Translation*, 30-31 March, Athens, Greece, pp.110-114.
- Duh, K. 2008. Ranking vs. regression in machine translation evaluation. *IN: Proceedings of the 3rd Workshop on Statistical Machine Translation*, 19 June, Columbus, Ohio, pp.191-194.
- Eisele, A. 2007. Hybrid machine translation: combining rule-based and statistical systems. *IN: Presented at the Conference of the First Machine Translation Marathon*, 16-20 April, Edinburgh, UK.
- Eisele, A. 2008. Hybrid architecture for machine translation. *IN: Presented at EuroMatrix the 2nd Machine Translation Marathon*, 14 May, Wandlitz, Germany.



- Elming, J. 2006. Transformation-based correction of rule-based MT. *IN: Proceedings of the 11th Annual Conference of the European Association for Machine Translation (EAMT2006)*, 19-20 June, Oslo, Norway, pp.219-226.
- Farghaly, A. 2003. *Handbook for Language Engineers*. Stanford, Calif.: CSLI Publication.
- Fauconnier, G. 1994. *Mental Spaces*. Cambridge: Cambridge University Press.
- Flanagan, Marian. 2009. *Recycling Texts: Human Evaluation of Example-Based Machine Translation Subtitles for DVD*. PhD thesis. Dublin City University.
- Flanagan, Mary A. 1994. Error Classification for MT Evaluation. *IN: Proceedings of the 1st Conference of the Association for Machine Translation in the Americas (AMTA 1994)*, 5-8 October, Columbia, Maryland, pp.65-72.
- Fleiss, J.L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*. 76(5), pp.378-382.
- Font Llitjós, A., Carbonell, J.G. and Lavie, A. 2005. A framework for interactive and automatic refinement of transfer-based machine translation. *IN: Proceedings of the 10th Annual Conference of the European Association for Machine Translation (EAMT)*, 30-31 May, Budapest, pp.30-31.
- Frey, L.R., Botan, C.H., Friedman, P.G. and Kreps, G.L. 1991. *Investigating Communication – An Introduction to Research Methods*. London: Prentice Hall International.
- Gerber, L. and Yang, J. 1997. Systran MT dictionary development. *IN: Proceedings of 5th Machine Translation Summit (MT Summit)*, San Diego, USA, pp. 211-217.
- Giménez, J.A. 2009. *Empirical Machine Translation and its Evaluation*. PhD thesis. Universitat Politècnica de Catalunya.
- Giménez, J.A., Amigó, E. and Hori, C. 2005. Machine translation evaluation inside QARLA. *IN: Proceedings of the International Workshop on Spoken Language Technology (IWSLT2005)*, 24-25 October, Pittsburgh, PA, USA, (no page number).
- González, J., Lagarda, A.L., Navarro, J.R., Eliodoro, L., Giménez, A., Casacuberta, F., De Val, J.M. and Fabregat, F. 2006. SisHiTra: A Spanish-to-Catalan hybrid

- machine translation system. *IN: Proceedings of the 5th SALT MIL Workshop on Minority Languages*, 23 May, Genoa, Italy, pp.69-73.
- Groves, D. and Way, A. 2005. Hybrid example-based SMT: the best of both worlds? *IN: Proceedings of the ACL 2005 Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, 29-30 June, Michigan, USA, pp.183-190.
- Gustavii, E. 2005. Target language preposition selection – an experiment with transformation-based learning and aligned bilingual data. *IN: Proceedings of the 10th Annual Conference of the European Association for Machine Translation – Practical Applications of Machine Translation*, 30-31 May, Budapest, Hungary, pp.112-118.
- Guzmán, R. 2008. Advanced automatic MT post-editing. *Multilingual*. 19(3), pp.52-57.
- Habash, N. 2002. Generation-heavy hybrid machine translation. *IN: Proceedings of the International Natural Language Generation Conference (NLG2002)*, 1-3 July, 2002, Harriman, New York, (no page number).
- Hartrumpf, S. 1999. Hybrid disambiguation of prepositional phrase attachment and interpretation. *IN: Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 21-22 June, University of Maryland, College Park, pp.111-120.
- Hartrumpf, S., Helbig, H. and Osswald, R. 2006. Semantic interpretation of prepositions for NLP applications. *IN: Proceedings of the 3rd ACL-SIGSEM Workshop on Prepositions*, 3 April, Trento, Italy, pp.26-36.
- Hatch, E.M. and Farhady, H. 1982. *Research Design and Statistics for Applied Linguistics*. Cambridge, Massachusetts, Newbury House.
- Hovy, E., King, M. and Popescu-Belis, A. 2002. Principles of context-based machine translation evaluation. *Machine Translation*. 17 (1), pp.43-75.
- Huijsen, W.O. 1998. Controlled language – an introduction. *IN: Proceedings of the 2nd Controlled Language Applications Workshop (CLAW1998)*, 21-22 May, Pittsburgh, Pennsylvania, pp.1-15.
- Hutchins, J. 2000. *Early Years in Machine Translation: Memoirs and Biographies of Pioneers*. Amsterdam: John Benjamins.

- Hutchins, J. and Somers, H. 1992. *An Introduction to Machine Translation*. London: Academic Press Limited.
- Isabelle, P., Goutte, C. and Simard, M. 2007. Domain adaptation of MT systems through automatic post-editing. *IN: Proceedings of the 11th Machine Translation Summit of the International Association for Machine Translation (MT Summit XI)*, 10-14 September, Copenhagen, Denmark, pp.255-261.
- Jurafsky, D. and Martin, J. 2009. Speech and language processing. *An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2nd edition. New Jersey/London: Pearson/Prentice Hall.
- Kennedy, G.D. 1998. *An Introduction to Corpus Linguistics*. London; New York, Longman.
- King, J.E. 2004. Software solutions for obtaining a kappa-type statistic for use with multiple raters. *IN: Presented at the Annual Meeting of the Southwest Educational Research Association*, 5-7 February, Dallas, Texas. [Online] Available from: <http://www.ccitonline.org/jking/homepage/interrater.html> [Accessed 5 September 2010].
- King, M., Popescu-Belis, A. and Hovy, E. 2003. FEMTI: creating and using a framework for the MT evaluation. *IN: Proceedings of the Machine Translation Summit IX*, 23-27 September, New Orleans, USA, pp.224-231.
- Knight, K. and Chander, I. 1994. Automated post-editing of documents. *IN: Proceedings of the 12th National Conference on Artificial Intelligence*, 31 July - 4 August, Seattle, Washington, pp.779-784.
- Knowles, K. 1978. Error analysis of Systran output – a suggested criterion for the “internal” evaluation of translation quality and a possible corrective for system design. *IN: Snell, B.M. (ed.) Translating and the Computer*. North-Holland Publishing Company, pp.109-133.
- Koehn, P. 2003. *Noun Phrase Translation*. PhD Thesis. University of South California.
- Koehn, P. 2004. Statistical significance tests for machine translation. *IN: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 25-26 July, Barcelona, Spain, pp.388-395.
- Koehn, P. 2010. *Statistical Machine Translation*. Cambridge University Press.

- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Dyer, C., Cowan, B., Shen, W., Moran, C. and Bojar, O. 2007. Moses: open source toolkit for statistical machine translation. *IN: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) (Demonstration Session)*, 23-30 June, Prague, Czech Republic, pp.177-180.
- Kraemer, C.H. and Thiemann, S. 1987. *How Many Subjects? Statistical Power Analysis in Research*. Sage Publications, Inc.
- Krings, H.P. 2001. *Repairing Texts: Empirical Investigations of Machine Translation Post-editing Processes*. Kent State University Press.
- Lagoudaki, E. 2006. Translation memories survey 2006: users' perceptions around TM usage. *IN: Proceedings of the International Conference Translating and the Computer*, 16-17 November, London. [Online] Available from: <http://www.atril.com/docs/tmsurvey.pdf> [Accessed 7 September 2010].
- Lakoff, G., and Johnson, M. 1980. *Metaphors We Live By*. Chicago: University of Chicago Press.
- Landis, J.R. and Koch, G.G. 1977. The measurement of observer agreement for categorical data. *Biometrics*. 33, pp.159-174.
- LDC. 2005. Linguistic Data Annotation Specification: Assessment of fluency and adequacy in translations. [Online] Available from: <http://www ldc.upenn.edu/Catalog/docs/LDC2003T17/TransAssess02.pdf> [Accessed 7 September 2010].
- Lee, R. and Renzetti, C. 1993. *The Problems of Researching Sensitive Topics*. Sage, Newbury Park.
- Li, H., Japkowicz, N. and Barrière, C. 2005. English to Chinese translation of prepositions. *IN: Kegal, B. and Lapalme, G. (eds). AI2005, LANI 3501*, pp.412-416.
- Li, N.C. and Thompson, A.S. 1981. *Mandarin Chinese: A Functional Reference Grammar*. University of California Press.
- Lin, C. and Och, F.J. 2004. ORANGE: A method for evaluating automatic evaluation metrics for machine translation. *IN: Proceedings of the 20th International Conference on Computational Linguistics (Coling 2004)*, 23-27 August, Geneva, Switzerland, pp.501-508.

- Litkowski, K. and Hargraves, O. 2005. The preposition project. *IN: Proceedings of the 2nd ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and Their Use in Computational Linguistic Formalisms and Applications*, 19-21 April, Colchester, UK, pp.171-179.
- Liu, D. and Gildea, D. 2005. Syntactic features for evaluation of machine translation. *IN: Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarisation*, 29-30 June, Ann Arbor, Michigan, pp.25-32.
- Liu, Danqing (刘丹青). 2002. Circumposition in Chinese (汉语中的框式介词). *Modern Linguistics (当代语言学)*. 1(4), (no page number).
- Lü, Y.J., Huang, J. and Liu, Q. 2007. Improving statistical machine translation performance by training data selection and optimization. *IN: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP)*, 28-30 June, Prague, Czech Republic, pp.343-350.
- Ma, Y.J. 2009. *Constrained Word Alignment Models for Statistical Machine Translation*. PhD Thesis. Dublin City University.
- Mamidi, R. 2004. Disambiguating prepositions for machine translation using lexical semantic resources. *IN: Proceedings of National Seminar on Theoretical and Applied Aspects of Lexical Semantics*, 27-29 February, Hyderabad, India.
- Megerdumian, K. 2003. Text mining, corpus building and testing. *IN: Farghaly, A. (ed). Handbook for Language Engineers*. CSLI Publications, pp.1-56.
- Melamed, I.D., Green, R. and Turian, J.P. 2003. Precision and recall of Machine Translation. *IN: Proceedings of Human Language Technology North American Chapter of the Association of Computational Linguistics Conference (NAACL 2003)*, 27 May-1 June, Edmonton, Canada, pp.61-63.
- Mellebeek, B., Owczarzak, K., Van Genabith, J. and Way, A. 2006. Multi-engine machine translation by recursive sentence decomposition. *IN: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, Visions for the Future of Machine Translation*, 8-12 August, Cambridge, USA, pp.110-118.
- Meyer, C.F. 2002. *English Corpus Linguistics: An Introduction*. Cambridge, UK; New York: Cambridge University Press.

- Mindt, D. and Weber, C. 1989. Prepositions in American and British English. *World Englishes*. 8(2), pp.229-238.
- Mitamura, T. 1999. Controlled language for multilingual machine translation. *IN: Proceedings of MT Summit VII – MT in the Great Translation Era*, 13-17 September, Kent Ridge Digital Labs, Singapore, pp.46-52.
- Nagao, M. 1984. A framework of a mechanical translation between Japanese and English by analogy principle. *IN: Elithorn, A. and Banerfi, R. (eds.) Artificial and Human Intelligence*. Amsterdam: North-Holland, pp.173-180.
- Nießen, S., Och, F.J., Leusch, G. and Ney, H. 2000. An evaluation tool for machine translation: fast evaluation for MT research. *IN: Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, 31 May-2 June, Athens, Greece, pp.39-45.
- Nomura, H. and Isahara, H. 1992. Evaluation surveys: the JEIDA methodology and survey. *IN: Proceedings of MT Evaluation: Basis for Future Directions*, 2-3 November, San Diego, California, pp.11-12.
- O'Brien, S. 2006. *Machine-Translatability and Post-Editing Effort: an Empirical Study Using Translog and Choice Network Analysis*. PhD thesis. Dublin City University.
- O'Brien, Sharon. 2003. Controlling controlled English: an analysis of several controlled language rules sets. *IN: Proceedings of EAMT-CLAW-03*, 15-17 May, Dublin, Ireland, pp.105-114.
- Och, F.J. 2003. Minimum error rate training for statistical machine translation. *IN: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL2003)*, 7-12 July, Sapporo, Japan, pp.160-167.
- Och, F.J. and Ney, H. 2003. A systematic comparison of various statistical alignment modes. *Computational Linguistics*. 29(1), pp.19-51.
- Olteanu, M. and Moldovan, D. 2005. PP-attachment disambiguation using large context. *IN: Proceedings of the Conference on Human Language Technology (HLT) and Empirical Methods in Natural Language Processing (EMNLP)*, 6-8 October, Vancouver, Canada, pp.273-280.
- Owczarzak, K., Van Genabith, J. and Way, A. 2007a. Dependency-based automatic evaluation for machine translation. *IN: Proceedings of the Workshop on Syntax and Structure in Statistical Machine Translation of HLT-NAACL*, 22-27 April, Rochester, NY, pp.86-93.

- Owczarzak, K., Graham, Y. and Van Genabith, J. 2007b. Using F-structures in machine translation evaluation. *IN: Proceedings of the LFG07 Conference*, 28-30 July, Stanford, California, pp.383-396.
- Ozdowska, S. and Way, A. 2009. Optimal bilingual data for French-English PB-SMT. *IN: Proceedings of the 13th Annual Meeting of the European Association for Machine Translation (EAMT 2009)*, 14-15 May, Barcelona, Spain, pp.96-103.
- Papineni, K., Roukos, S., Ward, T. and Zhu, W.J. 2002. BLEU: A method for automatic evaluation of machine translation. *IN: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL2002)*, 6-12 July, Philadelphia, PA, pp.311-318.
- Pierce, J.R., Carroll, J.B., Hamp, E.P., Hays, D.G., Hockett, C.F., Oettinger, A.G. and Perlis, A. 1966. Language and machines: computers in translation and linguistics. *Technical report of Automatic Language Processing Advisory Committee (ALPAC1966)*, National Academy of Sciences, National Research Council, Washington, DC, USA.
- Plitt, M. and Masselot, F. 2010. A productivity test of statistical machine translation post-editing in a typical localisation context. *IN: The Prague Bulletin of Mathematical Linguistics*. 93, January, pp.7-16.
- Pullum, G. and Huddleston, R. 2002. Prepositions and prepositional phrases. *IN: Huddleston and Pullum (eds.) The Cambridge Grammar of the English Language*. Cambridge University Press, pp.597-661.
- Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. 1985. *A Comprehensive Grammar of the English Language*. London and New York: Longman.
- Rapp, R. 2009. The back-translation score: automatic MT evaluation at the sentence level without reference translations. *IN: Proceedings of the Association for Computational Linguistics – International Joint Conference on Natural Language Processing 2009 Conference Short Papers*, 2-7 August, Suntec, Singapore, pp.133-136.
- Raybaud, S., Lavecchia, C., Langlois, D. and Smaili, K. 2009. Word- and sentence level confidence measures for machine translation. *IN: Proceedings of the 13th Annual Conference of the European Association for Machine Translation (EAMT 2009)*, 14-15 May, Barcelona, pp.104-111.
- Roturier, J. 2006. *An Investigation into the Impact of Controlled English Rules on the Comprehensibility, Usefulness and Acceptability of Machine-translated*

*Technical Documentation for French and German Users*. PhD Thesis. Dublin City University.

Roturier, J. 2009. Deploying novel MT technology to raise the bar for quality: A review of key advantages and challenges. *IN: Proceedings of the 12th Machine Translation Summit*, 26-30 August, Ottawa, Canada, pp.1-8.

Roturier, J. and Senellart, J. 2008. Automation of post-editing in localisation workflows. *IN: Presented at LISA Europe Forum 2008*, Dublin, Ireland.

Roturier, J., Krämer, S. and DÜchting, H. 2005. Machine translation: the translator's choice. *IN: Proceedings of the 10th Localisation Research Centre Conference (LRC X)*, 13-14 September, Limerick, Ireland.

Russo-Lassner, G., Lin, J. and Resnik, P. 2005. A paraphrase-based approach to machine translation evaluation. *Technical Report LAMP-TR-125/CS-TR-4754/UMIACS-TR-2005-57*, University of Maryland, College Park, Maryland.

Saint-Dizier, P. 2005. An overview of PrepNet: abstract notions, frames and inferential patterns. *IN: Proceedings of the 2nd ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, 19-21 April, Colchester, UK, pp.155-169.

Saint-Dizier, P. and Vazquez, G. 2001. A compositional framework for prepositions. *IN: Proceedings of the 4th International Workshop on Computational Semantics (IWCS-4)*, 10-12 January, Tilburg, Netherlands, pp.165-179.

Schwenk, H., Abdul-Rauf, S., Barrault, L. and Senellart, J. 2009. SMT and SPE machine translation systems for WMT'09. *IN: Proceedings of the 4th EACL Workshop on Statistical Machine Translation (WSMT2009)*, 30-31 March, Athens, Greece, pp.130-134.

Senellart, J. 2007. SYSTRAN MT/TM Integration. *ClientSide News Magazine*. [Online] Available from: <http://www.translationdirectory.com/article532.htm> [Accessed 7 September 2010].

Senellart, J., Simard, M. and Ueffing, N. 2010. Automatic post-editing. *Multilingual*. 21(2), pp.43-46.

Spethman, M., Rosas, M.L., Zhu, H. and Singh, N. 2009. *Website Globalization and E-business China*. [online] Available from:



<http://www.globalizationpartners.com/media/191397/china.pdf> [Accessed 7 September 2010].

- Simard, M., Goutte, C. and Isabelle, P. 2007a. Statistical phrase-based post-editing. *IN: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-HLT 2007)*, 22-27 April, Rochester, New York, pp.508-515.
- Simard, M., Ueffing, N., Isabelle, P. and Kuhn, R. 2007b. Rule-based translation with statistical phrase-based post-editing. *IN: Proceedings of the 2nd Workshop on Statistical Machine Translation*, 23 June, Prague, Czech Republic, pp.203-206.
- Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J. 2006. A study of translation edit rate with targeted human annotation. *IN: Proceedings of 7th Conference of the Association for Machine Translation in the Americas (AMTA2006)*, 8-12 August, Cambridge, USA, pp.223-231.
- Snover, M., Madnani, N., Dorr, B.J. and Schwartz, R. 2009. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. *IN: Proceedings of the EACL-2009 Workshop on Statistical Machine Translation (WMT09)*, 30-31 March, Athens, Greece, pp.259-268.
- Somers, H. 1999. Review article: Example-based machine translation. *Machine Translation*. 14(2), pp.113-157
- Somers, H. 2005. Round-trip translation: What is it good for? *IN: Proceedings of the Australasian Language Technology Workshop 2005*, December, Sydney, pp.127-133.
- Somers, H. and Fernández Díaz, G. 2004. Translation memory vs. example-based MT: What is the difference? *International Journal of Translation*. 16(2), pp.5-33.
- Sproat, R., Shih, C., Gale, W. and Chang, N. 1996. A stochastic finite-state word-segmentation algorithm for Chinese. *Computational Linguistics*. 22(3), pp.377-404.
- Stott, R. and Chapman, P. 2001. *Grammar and Writing*. Longman.
- Sun, Y.L. 2010. Mining the correlation between human and automatic evaluation. *IN: Proceedings of the 7th Language Resources and Evaluation Conference (LREC 2010)*, 19-21 May, Valetta, Malta, pp.1726-1730.

- Sun, Y.L., O'Brien, S., O'Hagan, M. and Hollowood, F. 2010. A novel pre-processing method for a rule-based machine translation system. *IN: Proceedings of the 14th Annual Conference of the European Association for Machine Translation (EAMT)*, 27-28 May, Saint-Raphaël, France.
- Surcin, S., Lange, E. and Senellart, J. 2007. Rapid development of new language pairs at SYSTRAN. *IN: Proceedings of the 11th Machine Translation Summit of the International Association for Machine Translation (MT Summit XI)*, 10-14 September, Copenhagen, Denmark, pp.443-449.
- Tatsumi, M. and Sun, Y.L. 2008. Linguistic comparison and analysis of statistical post-editing between Chinese and Japanese. *Journal of Localisation Focus*. 7(1), pp.22-33.
- TAUS report. 2009. *Taking the MT decision: selection, built-out and hosting*. [Online] Available from: <http://www.translationautomation.com/best-practices> [Accessed 1 May 2010].
- TAUS report. 2010. *Open MT: ready for business?* [Online] Available from: <http://www.translationautomation.com/technology> [Accessed 1 May 2010].
- Thurmair, G. 2005. Hybrid architectures for machine translation systems. *Language Resources and Evaluation*. 39(1), pp.91-108.
- Thurmair, G. 2009. Comparing different architectures of hybrid Machine Translation systems. *IN: Proceedings of MT Summit XII*, 26-30 August, Ottawa, Canada, pp.340-347.
- Turian, J.P., Shen, L. and Melamed, I.D. 2003. Evaluation of machine translation and its evaluation. *IN: Proceedings of the MT Summit IX*, 23-27 September, New Orleans, USA, pp.386-393.
- Van Rijsbergen, C. 1979. *Information Retrieval*. 2nd edition. London: Butterworths.
- Van Slype, G. 1979. Critical methods for evaluating the quality of machine translation. *Technical Report BR-19142* (Prepared for the European Commission Directorate General Scientific and Technical Information and Information Management. [Online] Available from: <http://www.issco.unige.ch/projects/isle> [Accessed 24 March 2010].
- Vandeghinste, V., Dirix, P., Schuurman, I., Markantonatou, S., Sofianopoulos, S., Vassiliou, M., Yannoutsou, O., Badia, T., Melero, M., Boleda, G., Carl, M. and Schmidt, P. 2008. Evaluation of a machine translation system for low resource languages: METIS-II. *IN: Proceedings of the 6th Edition of the Language*

*Resources and Evaluation Conference (LREC)*, 28-30 May, Marrakech, Morocco, pp.449-456.

Vauquois, B. 1968. A Survey of formal grammars and algorithms for recognition and transformation in machine translation. *IN: Proceedings of the International Federation for Information Processing (IFIP) Congress*, 5-10 August, Edinburgh, UK, pp.254-260.

Viera, A.J. and Garrett, J.M. 2005. Understanding interobserver agreement: the Kappa statistic. *Family Medicine*. 37(5), pp.360-363.

Vilar, D., Leusch, G., Ney, H. and Banchs, R.E. 2007. Human evaluation of machine translation through binary system comparisons. *IN: Proceedings of the 2nd Workshop on Statistical Machine Translation*, 23 June, Prague, Czech Republic, pp.96-103.

Vilar, D., Xu, J., D'Haro, L. and Ney, H. 2006. Error analysis of statistical machine translation output. *IN: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, 24-26 May, Genoa, Italy, pp.697-702.

Wang, C., Collins, M. and Koehn, P. 2007. Chinese syntactic reordering for statistical machine translation. *IN: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning*, 28-30 June, Prague, Czech Republic, pp.737-745.

Way, A. 2010. Machine translation. *IN: Clark, A., Fox, C. and Lappin, S. (eds.) The Handbook of Computational Linguistics and Natural Language Processing*. Wiley-Blackwell Press, pp.531-573.

White, J.S. 2003. How to evaluate machine translation. *IN: Somers, H. (ed.) Computers and Translation: A Translator's Guide*. Amsterdam/Philadelphia: John Benjamins Publishing, pp.211-244.

White, J.S., O'Connell, T. and O'Mara, F. 1994. The ARPA MT evaluation methodologies: evolution, lessons and further approaches. *IN: Proceedings of the 1st Conference of the Association for Machine Translation in the Americas*, 5-8 October, Columbia, Maryland, pp.193-205.

Woods, A., Fletcher, P. and Hughes, A. 1986. *Statistics in Language Studies*. New York: Cambridge University Press.

Wu, X.H., Cardey, S. and Greenfield, P. 2006. Some problems of prepositional phrases in machine translation. *IN: Proceedings of the 5th International*

- Conference on Natural Language Processing*, 23-25 August, Turku, Finland, pp.593-603.
- Xia, F. 2000. The segmentation guidelines for the Penn Chinese treebank (3.0). [Online] Available from: <http://www.cis.upenn.edu> [Accessed 16 June 2010].
- Xia, F. and McCord, M. 2004. Improving a statistical MT system with automatically learned rewrite patterns. IN: *Proceedings of the 20th international conference on Computational Linguistics*, 23-27 August, Geneva, Switzerland, pp.508-514.
- Xu, Jianping (许建平). 2003. *A practical course of English-Chinese and Chinese-English translation (英汉互译实践与技巧)*. 2nd edition. Tsinghua University Press.
- Xu, Y.S. and Seneff, S. 2008. Two-stage translation: a combined linguistic and statistical machine translation framework. IN: *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas (AMTA2008)*, 21-25 October, Hawaii, pp.222-231.
- Yang, J., Senellart, J. and Zajac, R. 2003. Systran's Chinese word segmentation. IN: *Proceedings of the 2nd Workshop of Special Interest Group on Chinese Language Processing (SIGHAN)*, 7-12 July, Sapporo, Japan.
- Yip, P.C. and Rimmington, D. 2004. *Chinese: A Comprehensive Grammar*. London and New York.
- Yu, Shiwen (俞士汶). 1994. Grammatical categorisation of modern Chinese (关于现代汉语的语法功能分类). *China Information World (中国计算机报)*. 2, pp.173-75.
- Zaidan, O. and Callison-Burch, C. 2009. Feasibility of human-in-the-loop minimum error rate training. IN: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6-7 August, Suntec, Singapore, pp.52-61.
- Zhu, Dexi (朱德熙). 2004. *Lecture Notes on Chinese Grammr (语法讲义)*. 2<sup>nd</sup> ed. Beijing: The Commercial Press.

## **Appendices**

## Appendix A – Questionnaire for the First Human Evaluation

### QUESTIONNAIRE

1- Are you a full-time or a part-time translator?

☐ Full-time

☐ Part-time

2- Please specify the rough number of words you have translated.

3- Have you ever taken courses on your native language grammar?

☐ At university

☐ In high-school

☐ No

4- Have you ever post-edited machine translation output professionally?

☐ Yes

☐ No

5- If you answered yes to question number 4, please specify the rough number of words you have post-edited.

6- Do you like working with machine translation output?

☐ 1- Not at all

☐ 2- Somewhat

☐ 3- Moderate

☐ 4- Very much

*Thank you very much for your participation!*

## Appendix B – Instructions for the First Human Evaluation

### 人工测评说明

#### 在开始人工测评前请仔细阅读下列说明。

此人工测评的主要目的在于统计在此类文档中，哪些英语**介词**在机器翻译中容易出错以及主要的错误类型有哪些。测评表中，每一句英文中只有一个介词短语被突显（介词加粗）。虽然多数情况下，您需要阅读整个英语句子及汉语翻译后才能对突显介词的翻译进行测评，但是请集中判断此**突显介词**的翻译。

该人工测评包含两项任务：

- 第一，请判断您是否需要**对介词或介词短语**的翻译进行后处理/编辑，“是” — 如果翻译不准确，例如存在语序错误或其他错误；“否” — 如果该翻译正确表达了原文意思且语序正确（尽管表达方式有点拗口）。从题为 “Need PE” 栏的下拉菜单中选择 “是” (Yes) 或 “否” (No) 。
- 第二，对需要进行后处理/编辑的句子，请判断介词或介词短语的翻译中包含哪些错误类型，从题为 “Error Category” 栏的下拉菜单中选择。开始正式测评前，请务必阅读测评表第一页中有关错误类型的样例。主要错误类型如下：
  1. **位置错误**：
    - i. **词级**：介词本身的翻译在汉语句子里处于错误的位置，需要调整；
    - ii. **短语级**：整个突显介词短语的翻译处于错误的位置，需要调整；
  2. **选词错误**：由于介词本身的多义性及对应汉语翻译的多样性，突显介词的翻译选词错误，需要替换。请不要受非介词翻译错误的影响，例如，如果介词短语中介词翻译正确，介词短语的位置正确，而介词短语中**名词短语翻译**有误，则不属于介词翻译的错误。
  3. **翻译不完整**：（有些汉语介词需要框式结构（例如：in: 在……中/方面/里面）以表达清楚题意）介词的翻译中缺少**必须**的后置介词；
  4. **翻译丢失**：没有对突显介词进行必要的翻译从而导致句意不清楚；

如果您觉得错误类型不在所选之列，请在 “Comment” 栏中简单描述。同时，欢迎在 “Comment” 栏中注明您的评论。如有其他问题，请随时和我联系。

邮件： sunsbecky@gmail.com

电话：00353-87-3141047

十分感谢您的耐心参与！

## Human Evaluation Instructions

### Please read the following instructions fully before commencing your evaluation.

The purpose of this evaluation is to see which English prepositions or prepositional phrases are incorrectly translated and what errors are displayed in their translation. In each English sentence, only **one** prepositional phrase is highlighted with the preposition in bold. While in most cases, you need to read the whole sentence and translation to come to a judgement, please focus on evaluating the **highlighted preposition or prepositional phrases**.

There are two tasks in this evaluation:

1. Please judge whether the translation of the **highlighted preposition (or prepositional phrase)** needs post-editing, i.e. “yes”- if the translation is not a good representation of the Source Text and there are errors in the translation; or “no”- if the translation is accurate (though perhaps not stylistically refined). Choose your answer (Yes or No) from the drop-down list in Column named “Need PE”.
2. If a translation needs post-editing, please choose the error categories from the drop-down lists in Columns named “Error Category”. Samples of error categories can be found in the following attachment as well as in Sheet 1 of the final evaluation sheet. Please look at the samples **before** commencing your evaluation. Several error categories are pre-defined as follows.

- **Incorrect position:**

- i. **Word level:** the translation of the bold preposition itself is at an incorrect position and should be moved to another position.
- ii. **Phrase level:** the translation of the whole highlighted prepositional phrase is at an incorrect position and should be reordered.

- **Incorrect lexical selection:** as English prepositions are polysemous and translations depend highly on context, the translation of the highlighted preposition is not correct for its context and should be changed to another translation. Please try not to let non-preposition related errors distract your judgment, for example, the highlighted prepositional phrase is at correct position and the translation of the bold preposition is correct, however, the translation of the noun phrase within the prepositional phrase is incorrect. This does not belong to preposition error.
- **Incomplete translation:** Sometimes, English prepositions must be translated into circumpositions in Chinese (e.g. in: 在.....中/方面/里面). The translation



for the highlighted preposition is not complete and the postposition should be added to make the translation correct.

- **Translation missing:** No translation for the highlighted preposition is found and should be added.

Please note, if no error category can describe the error you find, you may leave the error category blank, but please give a short description of the error in the Comment Column. Any other comments are warmly welcomed. Please put them also in the Comment Column. If you have any questions, please don't hesitate to contact me.

Email: [sunsbecky@gmail.com](mailto:sunsbecky@gmail.com); Phone: 00353-87-3141047

**Thank you very much for your participation!**

### 错误类型示例及解释 (Samples of Error Categories)

No need for post-editing		
Source	MT-output	Need PE?
In basic configuration, the LAN Enforcer performs the host authentication but is also configured to work with a RADIUS server, which provides user authentication.	在基本动词变位方面, LAN Enforcer 进行主机验证, 但是也配置与 RADIUS 服务器一起使用, 提供用户身份验证。	No
Note: Although the translation of "configuration" is incorrect here, it does not belong to preposition error.		

Incorrect position-word level			
Source	MT-output	Need PE	Error 1
About the status of infected computers.	受感染的计算机的状态关于。	Yes	Incorrect position - word level
Note: Translation of "About" should be moved to the front of the phrase. Correct translation should be: 有关受感染计算机的状态。			

Incorrect position-phrase level			
Source	MT-output	Need PE	Error
Add a warning to email messages about infected computers.	添加警告对关于受感染的计算机的电子邮件。	Yes	Incorrect position - phrase level
Note: "About..." should modify "warning" instead of "email message". Correct translation should be: 向电子邮件中添加有关受感染计算机的警告。Therefore, the translation of the highlighted prepositional phrase is at incorrect position.			

Incorrect lexical selection			
Source	MT-output	Need PE	Error 1
To add computers to the organizational unit and install software	添加计算机对组织单位和安装软件	Yes	Incorrect lexical selection
Note: Translation of "to" should be "到" instead of "对“			

Incomplete translation			
Source	MT-output	Need PE	Error 1
In the Security Status dialog box, review the features that trigger the Good and Poor status.	在安全状态对话框, 请查看触发好和恶劣的状态的功能。	Yes	Incomplete translation
Note: The translation of "in" is incomplete and the correct translation should be "在安全状态对话框中".			

Translation missing			
Source	MT-output	Need PE	Error 1
The client software creates file and folder scan exclusions for the following Microsoft Exchange server versions.	客户端软件创建文件和文件夹以下 Microsoft Exchange Server 版本的扫描排除。	Yes	Translation missing
Note: There is no translation for preposition “for”.			

## Appendix C – Instructions for the Second Human Evaluation

### 人工测评指南

#### 开始测评前，请详细阅读以下说明。

该人工测评的目的在于对比不同的翻译版本，从中找出最佳翻译。

本测评项目包含两项任务：

第一， 打开 “Sentence-Level Evaluation” 文件。需要测评的汉语翻译以黄色凸显，并标以 “Output 1/2/3/4”。不同版本的翻译顺序已经被打乱。此外，每句英文的汉语翻译数目并不相等。请仔细阅读英语原句、不同版本的翻译，然后对各个翻译从好到差进行排名（允许并列，比如，1 2 2 3）。提供的参考译文或许可以帮助您快速理解原文意思，**但是此参考译文并非“标准答案”，请以英文原意为准，只对比不同译文，并对译文进行排名。**从题为 “Ranking” 栏中选择排名。可以参考 Callison-Burch et al. (2008) 提出的以下标准进行评测：

- 流畅度：翻译的流畅自然程度；
- 准确度：翻译是否准确表达了原文的意思；

例如：

- 对有两个译文的句子：对比两个翻译，相对更流畅自然、翻译更准确的译文排第一名，选择 1；第二个译文选择 2；
- 对有三个译文的句子：对比三个翻译，最流畅、准确的译文，选择 1；相对比较流畅、准确的，选择 2；最差的选择 3；
- 对四个译文的句子：对比四个翻译，最流畅、准确的译文，选择 1；比较流畅、准确的，选择 2；勉强读懂的，选择 3；最拗口、不准确的翻译，选择 4；

第二， 完成第一个测评后，请打开第二个测评表 “Preposition Evaluation”。该测评表中的主要格式类似于第一个测评。不同之处在于每句中有一个介词或介词短语被凸显，阅读原文及对应翻译，对凸显的介词成分的翻译进行排名（允许并列）。请集中判断凸显的（红色）介词成分的翻译，不要受其他非介词翻译，例如分词不准确或者名词、动词翻译不对等问题的影响。从 “Ranking” 栏中选择排名。同样，参考译文仅供参考，并非“标准答案”，**不要将译文和参考译文进行对比；根据英文原意，在不同版本译文的内部对比并排序。**此外，个别句子中英汉凸现部分对应有可能不准确。

在两个测评任务中，如果你感觉两个或三个译文并没有质的区别，可以给予同样的排名。请在 “Comments” 栏中注明您的评论或意见。测评过程中如有任何问题，请随时和我联系。

邮件： sunsbecky@gmail.com

电话： 00353-87-3141047

十分感谢您的耐心参与！

## Human Evaluation Instructions

### **Please read the following instructions fully before commencing your evaluation.**

The purpose of this evaluation project is to compare different translations from different approaches and find out which translation is comparatively the best.

There are two tasks in this evaluation project.

1. Please open Sentence-Level Evaluation sheet. Translations from different approaches have been mixed, highlighted into yellow and marked by “Output 1/2/3/4”. The number of translations for each sentence is different. Please read the source sentence and all candidate translations, and then rank each candidate translation from best to worst relative to the other translations (ties are allowed, for example, 1 2 2 3). Reference translation is provided to help your understanding. However, you should always **rank the candidate translations comparing to the meaning of the source sentence instead of the reference as the reference is not the “Gold Standard”**. Select your rankings from the dropdown list in the Ranking Column following each translation. Two criteria could be used to rank different translations, namely Fluency and Accuracy. According to Callison-Burch et al (2008):

- Fluency: refers to how fluent the translation is;
- Accuracy: indicates how much of the meaning expressed in the source is also expressed in a translation.

For example,

- For sentences with two translations: Select 1 if the translation is comparatively more fluent and accurate than the other translation. Select 2 for the second translation;
- For sentences with three translations: Select 1 for the comparatively most fluent and accurate translation; 2 the second best translation in terms of fluency and accuracy; 3 for the last translation;
- For sentences with four translations: Select 1 for the most fluent and accurate translation; 2 for better translation; 3 for Ok translation and 4 for the least fluent and accurate translation;

2. After and only after you finish task 1, you can continue to the second task. Please open Preposition-evaluation sheet. The main format is similar to the first evaluation. Differently, in each sentence, one preposition or prepositional phrase has been highlighted into red. Please rank those constituent translation from best to worst relative to the other translations (ties are allowed). Please only grade the highlighted parts and don't be distracted by non-preposition errors such as

segmentation or noun/verb phrase mistranslation. Select your rank from the dropdown list named Ranking. The meaning of each rank is the same as in Task 1. Again reference is provided only to help your understanding; it is not “Gold Standard”. **Compare and rank each candidate translation based on the original meaning of the English sentence.** Please note that there might be errors in the process of highlighting, therefore, highlighted parts should be only taken as an appropriate guide. They might include extra words that not in the actual alignment, or miss words on either end.

As mentioned above, you can indicate a tie between two or more translations if there is no qualitative difference between the translations in either task. Any comments are warmly welcomed. Please put them in the Comment column. Please don’t hesitate to contact me if you have any question before or during the evaluation.

Email: [sunsbecky@gmail.com](mailto:sunsbecky@gmail.com)

Phone: 00353-87-3141047

Thank you very much for your participation!

## Appendix D –Questionnaire for the Second Human Evaluation

### QUESTIONNAIRE

1- Are you a full-time or a part-time translator?

☐ Full-time

☐ Part-time

2- If you answered Full-time for question 1, Please specify the rough number of years you have been a translator.

\_\_\_\_\_ years

3- Please specify the rough number of words you have translated.

\_\_\_\_\_ words

4- Have you ever post-edited machine translation output professionally?

☐ Yes

☐ No

5- How do you think of machine translation?

☐ 1- Very useful

☐ 2- Sometimes useful

☐ 3- Not useful

☐ 4- Useless

*Thank you very much for your participation!*

## Appendix E – Instructions for the Third Human Evaluation

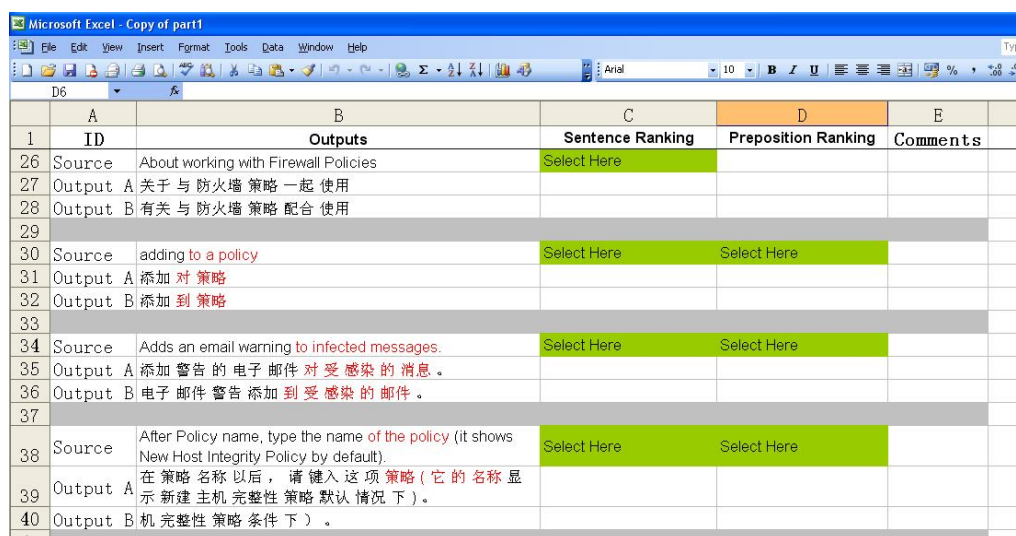
### 人工测评指南

开始测评前，请仔细阅读以下说明。由于原文件行数较多，现将其分为三个小文件。该说明对三个文件都适用。

该人工测评的目的在于对比不同的翻译版本并按照翻译质量进行排序。

对每一个英文句子，都有两个汉语翻译（Output A 和 Output B）。有些英文中有一个介词及其介词短语被凸显（介词短语用红色标示），其相应的汉语翻译也已经用红色标出。首先，请阅读英文原句及其两个汉语翻译，从整体上对比两个汉语翻译，然后判断**第一个翻译（Output A）更好，两个翻译质量等同，第二个翻译（Output B）更好**。在“Sentence Ranking”栏中标有“Select Here”的单元格的下拉菜单中选择您的判断。

然后，如果该句中沒有特别标红的介词短语，请继续下一句。如果该句中有介词短语标红，请集中阅读英文原句中标红的介词短语，然后对比其两个汉语翻译中的介词部分，判断**第一个翻译（Output A）更好，两个翻译质量等同，第二个翻译（Output B）更好**。判断时请将介词短语的**翻译**及其在句中的**位置**都考虑在内。在某些情况下，虽然介词的字面意思翻译正确，但在句中的位置却不正确，尤其是介词短语与相应动词的位置。在判断时，请集中判断介词本身的翻译和介词短语的位置。介词短语中实词（名词或动词）的翻译不应考虑在内。在“Preposition Ranking”栏中标有“Select Here”的单元格的下拉菜单中选择您的判断。（见下图）



	A	B	C	D	E
1	ID	Outputs	Sentence Ranking	Preposition Ranking	Comments
26	Source	About working with Firewall Policies	Select Here		
27	Output A	关于与防火墙策略一起使用			
28	Output B	有关与防火墙策略配合使用			
29					
30	Source	adding to a policy	Select Here	Select Here	
31	Output A	添加对策略			
32	Output B	添加到策略			
33					
34	Source	Adds an email warning to infected messages.	Select Here	Select Here	
35	Output A	添加警告的电子邮件对受感染的消息。			
36	Output B	电子邮件警告添加到受感染的邮件。			
37					
38	Source	After Policy name, type the name of the policy (it shows New Host Integrity Policy by default).	Select Here	Select Here	
39	Output A	在策略名称以后，请键入这项策略（它的名称显示新建主机完整性策略默认情况下）。			
40	Output B	机完整性策略条件下）。			
41					

请在“Comments”栏中注明您的评论或意见。测评过程中如有任何问题，请随时和我联系。

邮件：yanli.sun2@mail.dcu.ie

电话：00353-87-3141047

十分感谢您的参与！



### Human Evaluation Instructions

**Please read the following instructions fully before commencing your evaluation. Due to heavy formatting and the large number of lines, the evaluation sheet was split into three parts. This instruction is applicable to all the three sheets.**

The purpose of this evaluation project is to compare different translations and find out which translation is comparatively better.

For each English sentence (Source), there are two candidate Chinese translations (Output A and Output B). In some sentences, a prepositional phrase has been highlighted in red and in the other sentences no prepositional phrase has been highlighted.

Firstly, please read the source sentence and the two outputs, and then compare if **overall: Output A is Better** or **Output B is better** or the two translations are **Equal**. Please do not be distracted by the highlighted parts. Select from the dropdown list in the “Sentence Ranking” Column.

Secondly, if no prepositional phrase has been highlighted, please go to the next sentence. If there is a highlighted prepositional phrase in the sentence, please read the highlighted prepositional phrase and its two corresponding translations in Output A and B again. This time, **only** compare the two highlighted translations of the prepositional phrase. Select one of the following: **Output A is Better, Equal or Output B is Better** relative to the other translation. The **position**, as well as the **translation** of the highlighted preposition, should be taken into consideration as sometimes although the translation of one highlighted phrase is correct, the position is incorrect.

Please focus on the translation of the preposition and the position of the whole prepositional phrase. Translation of nouns or verbs in the prepositional phrases should not be taken into consideration. Select from the dropdown list (“Select Here”) in the “Preposition Ranking” column. (See the figure below)

Microsoft Excel - Copy of part1					
File Edit View Insert Format Tools Data Window Help					
D6					
	A	B	C	D	E
1	ID	Outputs	Sentence Ranking	Preposition Ranking	Comments
26	Source	About working with Firewall Policies	Select Here		
27	Output A	关于与防火墙策略一起使用			
28	Output B	有关与防火墙策略配合使用			
29					
30	Source	adding to a policy	Select Here	Select Here	
31	Output A	添加对策略			
32	Output B	添加到策略			
33					
34	Source	Adds an email warning to infected messages.	Select Here	Select Here	
35	Output A	添加警告的电子邮件对受感染的消息。			
36	Output B	电子邮件警告添加到受感染的邮件。			
37					
38	Source	After Policy name, type the name of the policy (it shows New Host Integrity Policy by default).	Select Here	Select Here	
39	Output A	在策略名称以后，请键入这项策略(它的名称显示新建主机完整性策略默认情况下)。			
40	Output B	机完整性策略条件下)。			
41					

Any comments are warmly welcomed. Please put them in the Comments column. Please don't hesitate to contact me if you have any question before or during the evaluation.

Email: [yanli.sun2@mail.dcu.ie](mailto:yanli.sun2@mail.dcu.ie)

Mobil Phone: 00353-87-3141047

Thank you very much for your participation!

## Appendix F –Questionnaire for the Third Human Evaluation

### QUESTIONNAIRE

1- Are you a full-time or a part-time translator?

☐ Full-time

☐ Part-time

2- If you answered Full-time for question 1, Please specify the rough number of years you have been a translator.  
years

3- Please specify the rough number of words you have translated or the average number of words you translate per day or per week?  
words in total Or words per day Or words per week

4- Have you ever post-edited machine translation output professionally?

☐ Yes

☐ No

5- How do you think of machine translation? Why?

☐ 1- Very useful

☐ 2- Sometimes useful

☐ 3- Not useful

☐ 4- Useless

Could you please put your reason here?

*Thank you very much for your participation!*