Syntax-based Pre-reordering for Chinese-to-Japanese Statistical Machine Translation

Author:
Dan HAN

Supervisor:
Yusuke MIYAO

DOCTOR OF PHILOSOPHY

2014

Dissertation submitted to the Department of Informatics,
School of Multidisciplinary Sciences,
The Graduate University for Advanced Studies (SOKENDAI)
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Advisory Committee

Assoc. Prof.	Yusuke MIYAO	National Institute of Informatics, SOKENDAI
Prof.	Akiko AIZAWA	National Institute of Informatics, TODAI
Prof.	Noriko KANDO	National Institute of Informatics, SOKENDAI
Assoc. Prof.	Nigel COLLIER	National Institute of Informatics, SOKENDAI
Prof.	Hideki ISOZAKI	Okayama Prefectural University

Abstract

Bilingual phrases are the main building blocks in statistical machine translation (SMT) systems. At training time, the most likely word-to-word alignment is computed and several heuristics are used to extract these bilingual phrases. Although this strategy performs relatively well when the source and target languages have a similar word order, the quality of extracted bilingual phrases diminishes when translating between languages structurally different, such as Chinese and Japanese. Syntax-based reordering methods in preprocessing stage have been developed and proved to be useful to aid the extraction of bilingual phrases and decoding. For Chinese-to-Japanese SMT, we carry out a detailed linguistic analysis on word order differences of this language pair to improve the word alignment. Our main contribution is threefold: (1) We first adapt an existing pre-reordering method called Head-finalization (HF) [1] for Chinese (HFC) [2] to improve Chinese-to-Japanese SMT system's translation quality. HF is originally designed to reorder English sentences for English-to-Japanese SMT and it performs well. However, our preliminary experiments results reveal its disadvantages on reordering Chinese due to particular characteristics of languages. We thus refine HF to HFC based on a deep linguistic study. To obtain the required syntactic information, we use a head-driven phrase structure grammar (HPSG) parser for Chinese. Nevertheless, the follow-up error analysis from the pre-reordering experiment explores more issues that bring difficulties for further improvement on HFC, such as the tree operation restriction of binary tree, inconsistency on definition of linguistic term and so on. (2) We then propose an entire new pre-reordering framework which is using an unlabeled dependency parser to achieve additional improvements on reordering Chinese sentences to be like Japanese word orders. We refer to it as DPC [3] for short. In this method, we first identify blocks of Chinese words that demand reorderings, such as verbs and certain particles. Then, we detect the proper position which is the right-hand side of their rightmost object dependent, since our reordering principle is to reorder a Subject-Verb-Object (SVO) language to resemble a Subject-Object-Verb (SOV) language. Other types of particles are relocated in the last step. Unlike other reordering systems, the boundaries of verbal blocks and their rightmost object in DPC are defined only by the dependency tree and part-of-speech tags.

Additionally, dismissing of using structural and punctuation border is another benefit for the reordering of the reported speech frequently occurring in news domain. The experiments show advantages of DPC over the SMT baseline (Moses) and our HFC systems. Important advantages of this method are the applicability of many reordering rules to other SVO and SOV language pairs as well as the availability of dependency parsers and POS-taggers for many languages. Considering our pre-reordering methods of HFC and DPC are linguistically-motivated, both are sensitive to parsing errors, even though DPC is designed to be more fault-tolerant parsing method by reducing the use of syntactic information, i.e., dependency labels. For future work on improving DPC or other reordering methods, it is meaningful to observe how parsing errors influence reordering performance. (3) We hence take a deep observation about the effects of parsing errors on reordering performance [4]. We combine empirical and descriptive approaches to carry out a three-stage incremental comparative analysis on the relationship between parsing and pre-reordering. Our conclusion can be used to benefit not only for the improvements of syntax-based pre-reordering methods, but also for the developments of POS taggers and parsers.

To my family

献给我的父亲和母亲

Acknowledgements

First and foremost I would like to express my sincere and deep gratitude to my advisor Professor Yusuke Miyao for his continuous support during my PhD studies. He is not only a talented researcher, but also a supportive supervisor. He patiently provided the vision, expertise, and encouragement for me to proceed through the whole doctoral program. Meanwhile, he has always given me great freedom to pursue independent work. Without his guidance and persistent help, this thesis would not have been possible. I truly appreciate all his contributions of time, ideas, and supervision of helping me through the rough road to become an independent researcher and making my PhD experience productive and stimulating.

I would also like to thank my committee members, Professor Akiko Aizawa, Professor Noriko Kando, Professor Nigel Collier, and Professor Hideki Isozaki, for their encouraging, insightful, and constructive comments and advice which guided me through all these years. Professor Aizawa has steadily provided priceless advice for my research ever since I started to prepare for the PhD entry examination. Professor Kando guided me in my study on information retrieval, and her enthusiasm on research was contagious for me. As female researchers, Professor Aizawa and Professor Kando have presented me the excellent examples as successful woman professors. I still remember all the valuable advice given by Professor Collier when I was taking the class "Presentation in English". It was a great experience which taught me how to make good presentations and communicate with people. The first paper that I read about machine translation is written by Professor Isozaki, and it becomes the guiding principle for my whole PhD research.

I am especially grateful for my first advisor in Japan, Professor Jun'ichi Tsujii, a great computer scientist who opened the door for me to enter the field of natural language processing. I have been given a unique opportunity to encounter so many remarkable individuals in Tsujii lab and it has been a source of friendships for me. I am deeply indebted to my former lab-mate, Dr. Xianchao Wu, for his generous help and teaching. He encouraged me the most when I was in a difficult time in the beginning of being a computer science student, and he introduced me to statistical machine translation, one of the most interesting and challenging research topics. Thanks to him I had the first opportunity to do an internship at NTT Communication Science Laboratories where I had the chance to learn from strong MT researchers, such as Dr. Masaaki Nagata, Dr. Katsuhito Sudoh, Dr. Kevin Duh and others. I would like to especially acknowledge Dr. Sudoh who was patiently guiding me walking in the field of MT during my stays at NTT.

It was the most valuable research experience for me of working in such an emphatic MT team, and most of my PhD work has been done at NTT.

I will forever be thankful to my lab-mate Dr. Pascual Martínez-Gómez. He is the primary resource that my science questions get answers from and always ready for helping me to crank out the difficulties throughout my PhD life. He consistently push and motivate me during the hard moments. It would take much longer for me to convert from an electronic engineer to a computer scientist without him. I also want to thank Professor Takuya Matsuzaki, Dr. Jun Hatori, Dr. Hiroki Hanaoka, for their support of my research, and valuable, thoughtful advices. All members from both Tsujii lab and Miyao lab, also deserve my sincerest thanks, Goran Topic, Luke McCrohon, Pontus Stenetorp, Sho Hoshino, Hiroshi Noji, Keiko Kigoshi, and so many others that I can not list all names here, their friendship and assistance has meant a lot to me.

My time at National Institute of Informatics was made enjoyable largely due to my dear friends, Lihua Zhao, Yunlong Feng, and many others. Special thanks to Lihua, who is an enthusiastic, considerate, and constructive friend. I remember that several times Lihua and I went for a coffee break from writing papers at 3 am in the morning in the lab.

Lastly, I thank the educational and financial support from National Institute of Informatics (NII) and the Graduate University for Advanced Studies (SOKENDAI) for my doctoral studies. I appreciate very much the platform that is built by the cooperation between Chinese government and Japanese government. I wish the best for the peace and development in these two countries and the world.

Lastly and most importantly, none of this would have been possible without the support of my family. I would like to express my heart-felt gratitude to my father, my mother, my grandmother, and my sister, who have been a constant source of unconditional love, concern, encouragement, and patience all these years. Their love provided my inspiration and was my driving force. I owe them everything and wish I could show them just how much I love and appreciate them. I hope that I makes them proud. I would like to thank them in Chinese:

谢谢你们对我无私奉献的爱和永不间断的支持，我爱你们。

Contents

Abstract **ii**

List of Figures **ix**

List of Tables **x**

1 Introduction **1**
- 1.1 Motivation . . . 3
- 1.2 Problem Statement . . . 4
- 1.3 Contributions . . . 6
- 1.4 Outline . . . 8

2 Background **10**
- 2.1 Statistical Machine Translation . . . 11
  - 2.1.1 Historical Perspectives . . . 11
  - 2.1.2 The Statistical Model . . . 12
  - 2.1.3 The Word Alignment Problem . . . 15
  - 2.1.4 Phrase Extraction . . . 17
- 2.2 Parsing . . . 19
- 2.3 Resource . . . 20

3 Related Work **24**
- 3.1 Language Independent Reordering . . . 25
- 3.2 Language Dependent Reordering . . . 26
- 3.3 Error Analysis . . . 28
- 3.4 Head Finalization (HF) . . . 29
- 3.5 Summary . . . 31

4 Head Finalization for Chinese (HFC) **32**
- 4.1 Preliminary adaptation of HF for Chinese . . . 33
- 4.2 Discrepancies in Head Definition . . . 35
  - 4.2.1 Aspect particle . . . 35
  - 4.2.2 Adverbial modifier bu4(not) . . . 36
  - 4.2.3 Sentence-final particle . . . 36
  - 4.2.4 *Et cetera* . . . 37

4.2.5 Head finalization for Chinese (HFC) 39
4.3 Evaluation 40
4.3.1 Experiment setting 40
4.3.2 Results 41
4.4 Error Analysis 44
4.4.1 Serial verbs 44
4.4.2 Complementizer 46
4.4.3 Adverbial modifier 47
4.4.4 Verbal nominalization and nounal verbalization 48
4.5 Summary 50

5 Dependency Parsing based Pre-reordering for Chinese (DPC) 51
5.1 Methodology 52
5.1.1 Identifying verbal block (Vb) 52
5.1.2 Identifying the right-most object dependent (RM-D) 57
5.1.3 Identifying other particles (Oth-DEP) 59
5.1.4 Summary of the reordering framework 60
5.2 Evaluation 63
5.2.1 Experiment setting 63
5.2.2 Results 63
5.3 Summary 70

6 Effects of Parsing Errors on Pre-reordering 71
6.1 Analysis Method 72
6.2 Preliminary Experiment 73
6.2.1 Gold Data 73
6.2.2 Evaluation 74
6.3 Analysis on Cause of Reordering Errors 80
6.3.1 Dependency parse errors by part-of-speech 82
6.3.2 Dependency parse errors by dependency structure 85
6.3.3 Further Analysis Possibilities 90
6.4 Summary 92

7 Final Remarks and Future Work 93
7.1 Discussion 94
7.2 Future Work 98
7.3 Conclusion 99

Appendices 100

A Summary of Part-of-Speech Tag Set in Penn Chinese Treebank 100

B Head Rules for Penn2Malt to Convert the Penn Chinese Treebank 102

Bibliography 104

List of Figures

2.1 Bernard Vauquois' pyramid 12
2.2 Statistical machine translation system 13
2.3 A simple example of decomposed translation generation process 14
2.4 Alignment comparison 16
2.5 Phrase extraction 17
2.6 After pre-reordering 19
2.7 An HPSG parse tree of Chinese 21
2.8 An unlabeled dependency parse tree 21

3.1 Head finalization (HF) 30

4.1 Simple HF adaptation for Chinese 34
4.2 Example for adverbial modifier (not) 37
4.3 Example for sentence-final particle 38
4.4 Example for *Et cetera* 38
4.5 Example for serial verb construction 45
4.6 Example for serial verbs in subordinate relationship 46
4.7 Example for complementizer 47
4.8 Example for adverbial modifier 49

5.1 Example of bei-construction 54
5.2 Detect and reorder a verbal block (Vb) 56
5.3 Coordination verb phrases 58
5.4 Reported speech 59
5.5 Example of dependency parsing based Chinese pre-reordering 62

6.1 The distribution of τ for 491 sentence pairs 77
6.2 The distribution of τ for $2,164$ sentence pairs 79
6.3 Example for calculating parsing errors in terms of POS tag. 83
6.4 The distribution of top three dependent-error POS tags and their tendency lines. 84
6.5 The distribution of top two head-error POS tags and their tendency lines. 84
6.6 Example for parsing error patterns of Root-A and RM_D-D 87
6.7 Example for parsing error patterns of Root-G and RM_D-C 88
6.8 Distribution of three types of dependency parsing errors 90
6.9 Distribution of different patterns of ROOT error 90
6.10 Distribution of different patterns of RM-D error 91

List of Tables

2.1 Statistical Characteristics of Corpora 23

4.1 The List of POS tags for Exception Reordering Rules 39
4.2 Evaluation Results of Translation Quality on News Domain 42
4.3 Evaluation Results of Translation Quality on Patent Domain 42
4.4 Reordering Examples of HF and HFC 43

5.1 List of POS tag Candidates 53
5.2 Reordering Examples of HFC and DPC 65
5.3 Evaluation of Translation Quality on News Domain (Training 1) 66
5.4 Evaluation of Translation Quality on News Domain (Training 2) 67
5.5 Evaluation of Translation Quality on Patent Domain (Training 1) 68
5.6 Evaluation of Translation Quality on Patent Domain (Training 2) 69

6.1 Statistics of Selected Sentences in Five Genres of CTB-7. 74
6.2 The distribution of τ for 491 sentence pairs 77
6.3 Sentence numbers of monotonic alignment in 491 sentence pairs 78
6.4 The distribution of τ for $2,164$ sentence pairs 79
6.5 Sentence numbers of monotonic alignment in $2,164$ sentence pairs 80
6.6 Reordering Examples of Gold-HFC and Auto-HFC 81
6.7 Reordering Examples of Gold-DPC and Auto-DPC 81
6.8 Dependency Error Patterns 86

A.1 POS tags defined in Penn Chinese Treebank 101

B.1 Head Rules for Converting Penn Trees 103

Chapter 1

Introduction

Translation between Chinese and Japanese languages gains interest as their economic and political relationship intensifies. Despite their linguistic influences, these languages have different syntactic structures. Linguistically, similar as English, Chinese is also known as Subject-Verb-Object (SVO) language or head-initial language, while Japanese is a typical Subject-Object-Verb (SOV) language, namely head-final language.

In state-of-the-art Statistical Machine Translation (SMT) systems, bilingual phrases are the main building blocks for constructing a translation given a sentence from a source language. To extract those bilingual phrases from a parallel corpus, the first step is to discover the implicit word-to-word correspondences between bilingual sentences [5]. Then, a symmetrization matrix is built [6] by using word-to-word alignments, and a wide variety of heuristics can be used to extract the bilingual phrases [7, 8]. This method performs relatively well when the source and the target languages have similar word order, as in the case of French, Spanish, and English since their sentences are all following the same pattern, namely Subject-Verb-Object (SVO). However, when translating between languages with very different sentence structures, as in the case of between English (SVO) and Japanese (SOV), or Chinese (SVO) and Japanese (SOV), the quality of extracted bilingual phrases and the overall translation quality diminish.

Current word alignment models [9] in phrase based SMT systems account for local differences in word order between bilingual sentences, but fail at capturing long distance word alignments. One of the main problems in the search of the best word alignment is the combinatorial explosion of word orders. In other words, the main reason behind the drop in translation quality is the difficulty to find word-to-word alignments between words of sentences from such language pairs. Methods that address local non-monotonic word alignments proved ineffective, since words in SVO sentences usually should align to words in SOV languages that are in a very different position in the sentence.

Traditional reordering models, i.e., lexicalized reordering models, operate during training and decoding. Estimating the likelihood of all possible word-to-word alignments and reordering possibilities in a bilingual sentence pair is a combinatorial problem, and its complete exploration is unfeasible for medium size sentence lengths. As it was introduced in this thesis, using knowledge of structural differences between SVO and SOV languages in a preprocessing stage, can not only reduce the huge computational cost for reordering and contribute to improve word-to-word alignments, but also reduce the constraints that are introduced by existing reordering models. A popular approach is to extract the syntactic structure of sentences from the source language, and reorder the source words to imitate the word order of the target language. This strategy is called pre-reordering, and parsers play an important role to extract the syntactic structure of source sentences.

There have been important advances in syntactic parsers, and different types of parsing technologies have been developed for languages, such as English and Chinese. In general, two types of parsers have been used to pre-reorder sentences in SMT, namely parsers following the paradigm on head-driven phrase structure grammar and dependency grammar. Both types of parsers infer the structure of sentences, but they are able to recover different information from the sentence, such as phrase constituents or dependency relations between words.

Despite of these considerable advances in parsing technology, current parsers are still not perfect and may produce some errors with a certain frequency. The issue of parsing errors in syntax-based pre-reordering is crucial, as it potentially affects the performance of reordering methods and impact the overall translation quality. Although there has been extensive research in pre-reordering methods for statistical machine translation, little attention has been paid to the influence of parsing errors to pre-reordering performance and overall machine translation quality.

1.1 Motivation

Textual content is constantly produced and shared in a globalized environment. Companies, governmental agencies and individuals are in constant need of translation services to make their content more accessible to international audiences. Professional human translators are capable of providing high quality translation services, but those services are often expensive and slow. After two decades of intense research on statistical machine translation, machines became capable of producing fast and inexpensive on-demand translations. Although machine translated text may satisfy some basic communication purposes, it is far from being acceptable in many domains such as legal, patent or news domains.

China and Japan have a long history of economical and political relations, but the machine translation community has not dedicated a major interest to the Chinese and Japanese language pair. Despite of the individual importance of those two languages, there was a significant scarcity of parallel corpora, which is essential to build and evaluate state-of-the-art machine translation systems. At the same time, data scarcity prevents other researchers from working on this language pair, closing an unfortunate vicious cycle.

There has been much investigation on machine translation between language pairs with similar word order, such as French and Spanish, or English and Chinese. However, many machine translation methods do not perform well when languages have different word orders. In spite of its many similarities, Chinese and Japanese have very different sentence structures, which poses an interesting challenge to current machine translation paradigms. For this reason, Chinese and Japanese language pair is a relevant case of a language pair with different word order for which new methods have to be devised.

In this study, we pursue a greater understanding of the relationship of word orders between Japanese and Chinese. Language word orders are governed by an underlying syntactic theory for that specific language, but relationships between syntactic theories of two languages remain unclear. We hope that the understanding of these relationships may lead us to improve automatic translation between Chinese and Japanese language pair. We believe that the present investigation will inspire other similar studies that need to tackle machine translation between language pairs with very different sentence structures, and to this purpose we dedicate the efforts in this thesis.

1.2 Problem Statement

Statistical Machine Translation (SMT) systems work in two stages. The first stage is the training of the system, where parameters of a set of models are estimated from data sets. The second stage (decoding) uses the estimated model parameters to translate sentences from a source language into sentences of a target language. In the first stage, there are two types of models that are of special interest in this study. The first type of models is translation models, which contain information on the candidate words of a target language that could be a translation of a set of words from the sentence in the source language. The second type of models are reordering models, which inform the decoding stage about the appropriate word order of the candidate words. Both types of models rely on the correct recognition of word-to-word correspondences between the bilingual sentences of the training data sets, the so called parallel text.

These bilingual sentences in the parallel text do not contain explicit information on how words from sentences in the source language correspond to words from sentences in the target language. For this reason, the SMT community uses unsupervised alignment methods to infer these word-to-word correspondences. In theory, these unsupervised alignment methods would have to explore all possible word-to-word correspondences to estimate their likelihood. In practice, however, exploring all possible word-to-word correspondences is a combinatorial problem that becomes intractable even for medium-size sentences. For this reason, unsupervised estimators of word-to-word correspondence usually explore only a subset of all possible combinations. This subset often corresponds to words from the sentence of the target language that are in a similar position to a given word from the sentence of the source language.

Although this is a suboptimal solution to the practical problem of exhaustive exploration of possible word alignments, it performs reasonably well in most popular language pairs, which share a similar sentence structure and word order. However, this sub-optimal solution does not perform well for sentences from language pairs with very different sentence structure such as Chinese and Japanese, since words in a Chinese sentence and their corresponding translations in the Japanese sentence may have very different word positions.

To alleviate this problem, different techniques have been proposed in the literature and they will be reviewed in Chapter 3. In our study, we focused in one of these techniques called pre-reordering, where words in Chinese sentences (source language) are re-arranged to resemble the word order of Japanese sentences (target language). Such pre-reordering operation is performed at a pre-processing stage on Chinese sentences of the parallel text before the training stage takes place. In the decoding stage, Chinese sentences have to be translated into Japanese sentences, and these Chinese sentences will be pre-reordered before the decoding occurs.

In order to automatically re-arrange words in Chinese sentences to resemble the word order of their Japanese counterparts, we need first to perform an analysis on word order differences between sentences of both languages with the objective to capture regularities in these order differences. These regularities in word order differences can then be transformed into reordering rules to re-arrange words in Chinese sentences to resemble word order in Japanese.

The first problem to solve will be to discover, analyze and characterize word order differences in terms of relevant linguistic features, with the objective to capture patterns in word order differences between both languages. The second problem will consist in expressing these patterns in word order differences in a usable manner, that is, to design reordering rules that preserve the meaning of the original sentence but that complies with the word order of the target language.

We work under the assumption that similar word orders between Chinese and Japanese sentences ease the recognition of word-to-word correspondences in training and decoding stages. Thus, we believe that our pre-reordering methods will improve overall machine translation quality. We will measure our level of success using two strategies. The first strategy will be to measure the similarity of word orders between reordered Chinese sentences and original Japanese sentences. The second strategy will be to measure the overall impact of our pre-reordering techniques in terms of translation quality.

1.3 Contributions

Studying differences in word order between Chinese and Japanese sentences is a challenging task due to the combinatorial nature of word ordering. For this reason, we study word order differences in terms of differences in the syntactic structure of Chinese and Japanese sentences. There are two influential trends in linguistic theory regarding to syntactic structure that have been adopted in the community of computational linguistics.

The first class of syntactic theory is phrase structure grammars or constituency grammars. The structure of this type has been studied in computer science and natural language processing by using head-driven phrase structure grammar (HPSG) [10]. In a constituency tree, every node contains syntactic and semantic information about the sub-constituent it represents. Within this information, there is *head* information which indicates what word from the constituent is the head of the phrase and will play an important role in part of our work. Another common class of syntactic theory is dependency grammar in which the syntactic structure in a sentence is represented using dependency relations between words. Unlike constituency structures, dependency structures are flatter and they lack a finite verb phrase constituent, which makes them suitable to analyze sentences from free word order languages (like Chinese) but at a higher expense in the complexity of the analysis result.

The contribution of this thesis is three-fold. We first analyze patterns of word order differences between Chinese and Japanese sentences in terms of HPSG structure. Isozaki et al. [1] noted that an important difference between Japanese and English sentences is the head position of phrases, and used such insight to translate English sentences to Japanese by moving heads of English phrases to the end of their constituent at a pre-processing stage, which is called Head Finalization (HF) pre-reordering method. While Japanese is a head-final language, English and Chinese are head-initial languages. For this reason, we re-implemented this technique to translate Chinese to Japanese, with little success due to discrepancies in the definition of *head* between Chinese and Japanese. To overcome this problem, we characterize those discrepancies in terms of part-of-speech (POS) tags of the words in the constituents. The key result of this analysis was a refinement of HF that we called Head Finalization for Chinese (HFC), and evaluated its pre-reordering performance in terms of well known machine translation quality metrics.

As for our second contribution, based on the findings of our work on HFC, we discarded the phrase structure grammars by reason of the tight tree structure. Alternatively, we analyzed word order differences in terms of differences in the dependency structure of Chinese and Japanese sentences and POS tags of their words. We found a wide range of patterns that characterize these word order differences and formulated them in a usable manner. The key result of this characterization was thus an entirely original pre-reordering method called Unlabeled Dependency Parsing based Pre-reordering for Chinese (DPC), which was evaluated again in terms of machine translation quality, displaying significant performance increase with respect to our baselines.

There are two main components in our machine translation pipeline that may affect overall translation quality. The first one is the performance of our proposed pre-reordering methods in reordering words of Chinese sentences to resemble word order of Japanese sentences. The second one is the performance of automatic parsers that are used to recognize phrase or dependency structure of Chinese sentences. Since our methods are based on these automatic extracted syntactic information to carry out the pre-reordering, parsing errors in the recognition of the sentence structure will affect pre-reordering performance and, ultimately, overall translation quality. Studying the effect of parsing errors on pre-reordering performance is useful to discover what types of the parsing errors are with negative consequences on pre-reordering and to gauge their impact in pre-reordering performance and translation quality.

Therefore, in our third contribution, we characterize such parsing errors for both types of syntactic structures. We expect that both the methodology and evidencing these patterns of parsing errors will help us and other researchers to design more robust pre-reordering methods.

1.4 Outline

This thesis consists of 7 chapters. We introduce the remaining chapters as follows:

- **Chapter 2: Background**
  We describe the basics of statistical machine translation that are relevant to our work, and what is the motivation behind pre-reordering to improve translation. Then, we do a small overview of parsing technologies for Chinese, and describe the corpora that have been available prior and during our work.

- **Chapter 3: Related Work**
  We introduce the general word ordering problem in statistical machine translation and describe what are the approaches that have been followed so far for different language pairs. We will make special emphasis on pre-reordering, its different techniques, and how it is expected to benefit machine translation. In the last section, we introduce in detail the philosophy of Head Finalization (HF) pre-reordering method that was initially developed to reorder words in English sentences to resemble the word order of Japanese sentences.

- **Chapter 4: Head Finalization for Chinese (HFC)**
  We first implement Head Finalization into Chinese and exhibit the result of our preliminary experiment. Then, we will explain its limitations when operated to reorder words in Chinese sentences. We present our linguistic analysis of discrepancies of head definitions between Japanese and Chinese, and present a refined method for Chinese that attempts to solve those discrepancies. We close this chapter by evaluating Head Finalization for Chinese (HFC) in terms of translation quality, and carry out an error analysis to evidence the limitations of this method.

- **Chapter 5: Unlabeled Dependency Parsing based Pre-reordering for Chinese (DPC)**

  In this chapter, we carry out an analysis of ordering differences between Japanese and Chinese in terms of their dependency structures and POS tags, as we think this is a minimal set of highly descriptive features to inform the pre-reordering. First, we identify structure differences between Japanese and Chinese sentences using dependency relations. Then, we identified POS tags that are strong signals to guide reordering. We devise rules to move sentence components from Chinese sentences to resemble word order of Japanese sentences. Finally, we evaluate the effectiveness of this method in terms of translation quality, and compare it with our Refined Head Finalization and other state-of-the-art baselines.

- **Chapter 6: Effects of Parsing Errors on Pre-reordering**

  We dedicate this chapter to the objective of characterizing and quantify parsing errors that affect negatively to pre-reordering performance of our methods. We carry out a descriptive and quantitative analysis of the impact of parsing errors on pre-reordering, by using manually and automatically parsed and reordered Chinese sentences.

- **Chapter 7: Conclusion**

  In this last chapter, we discuss the advantages and disadvantages of our pre-reordering methods that use phrase-based and dependency syntactic structures. Then, we elaborate our conclusions on the relationship between parsing errors and the pre-reordering performance of our methods, and point to potential applications that would benefit from our findings.

Chapter 2

Background

The statistical machine translation community has developed methods to obtain useful translations in some domains. There have been many excellent tutorial introductions [11–13] and a well-written book by Koehn [14] on SMT. In this chapter, we show a brief introduction of modern statistical machine translation techniques of which we mainly focus on two of them, word alignment and phrase extraction, since they are the most relevant to our work. Linguistically motivated heuristics are proved to be useful of guiding reordering and thereupon improving the word alignment between distance language pairs. Parsers are thus used to extract the required syntactic information. A compressed depiction of parsers for Chinese that we use for our work and a short description of the corpora that we use for evaluation will be given as well in two sections of this chapter.

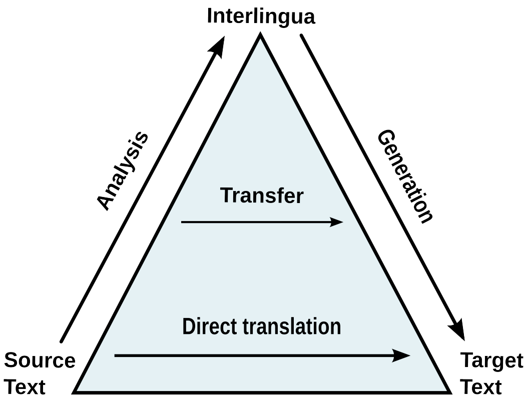
2.1 Statistical Machine Translation

2.1.1 Historical Perspectives

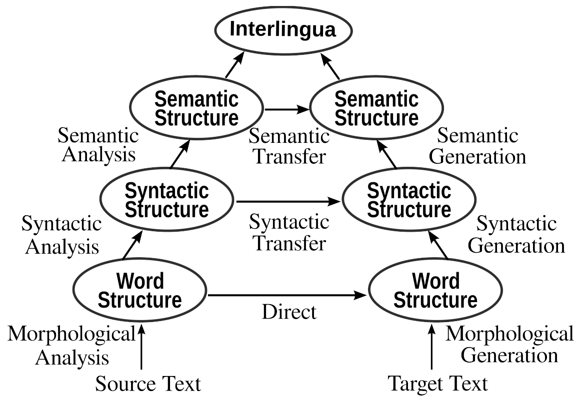
Since Dr. Warren Weaver first mentioned the idea of using computers to translate documents between natural human languages in a letter in 1947, which was later formed as a memorandum [15] in 1949, machine translation (MT) has drawn numerous researchers' attention throughout the world over about seven decades by now. Authors in [16, 17] had exhibited comprehensive historical overviews and surveys for machine translation. As can be seen from the past, research on MT has never stop completely, although once the ALPAC report in 1966 disappointed the over-optimistic MT research community [18], which resulted in funding loss almost entirely.

Before 1990s, rule-based approach was dominant that various rules were designed for syntactic analysis, lexical transfer, morphology and so on [19–22]. Three types of models were mainly explored in the early days, which are the simple direct translation model, the more sophisticated transfer model, and the interlingua model. These models are inspired by analyzing how languages are formed. Bernard Vauquois has drawn a famous pyramid diagram [23] (Figure 2.1a) for showing these MT systems' architectures (Figure 2.1b is taken from [16]).

Since 1989, new methods and strategies were proposed given the availability of reasonable amounts of human translations, which are roughly known as corpus-based approach. The emergence of such method had broken the monopoly status of rule-based approach. As one of the major directions in corpus-based studies, example-based machine translation (EBMT) used the idea of translation by analogy. Although it was first proposed by Nagao in 1981 [24], a flood of experiments started from the end of the 1980s [25–28]. Meanwhile, another direction of such empirical approach, which is known as statistical machine translation (SMT), has been re-introduced into the community [5, 29–31]. Researchers from IBM was motivated by the successes of statistical methods in speech recognition, and modeled the machine translation task as a machine learning optimization problem.



(A) Bernard Vauquois' pyramid



(B) Analysis in Bernard Vauquois' pyramid

FIGURE 2.1: Bernard Vauquois' pyramid presenting different types of MT systems.

Nowadays, as the rapid advances in computational power, open online toolkits, available text resources, etc., a number of factors contributed to the development of using statistical method on MT. The continuing improvements of the translation performance raise the hope again on machine translation without blind optimism. Currently, not only for the academic research purpose, but also for commercial products, many useful SMT systems has been developed, and the best performing ones are phrase-based. Since our work is based on a phrase-based SMT system, for the rest of Section 2.1, we will lay out the state-of-the-art statistical modeling methods for word alignment and phrase extraction, which are the most related parts to our interest in a phrase-based SMT system.

2.1.2 The Statistical Model

As the most investigated approach to machine translation, SMT uses machine learning methods to solve the natural language translation problem, which starts from large human-produced translation corpora. By observing a large number of high quality translation samples, SMT systems learn to automatically translate phrases and sentences with the highest probability from the source language to the target language. Figure 2.2 shows how a basic SMT system works. (Modified from the tutorial [32])

In the statistical model, in a source-target sentence pair $< \mathbf{f}, \mathbf{e} >$, sentences are defined as sequences of words, $\mathbf{f} = f_1, \ldots, f_i, \ldots, f_L$ represents the source sentence, while $\mathbf{e} = e_1, \ldots, e_j, \ldots, e_M$ represents the target sentence. When the system receive a source

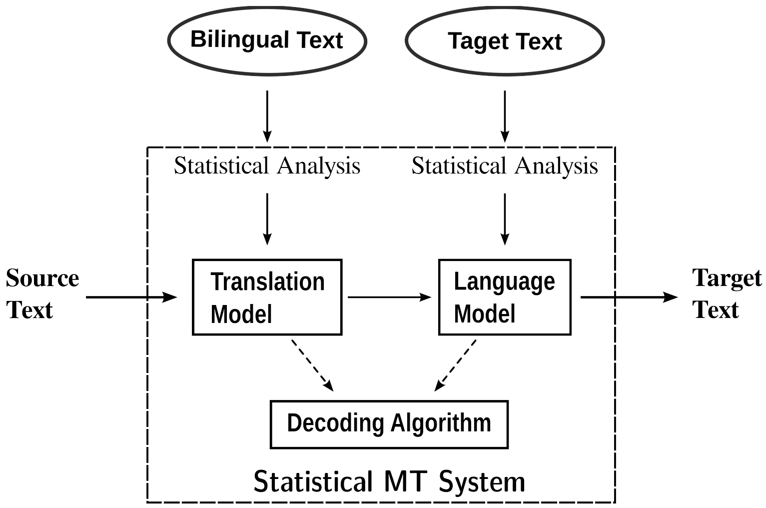


FIGURE 2.2: A basic architecture of a statistical machine translation system.

sentence $\mathbf{f}$, there are many translation possibilities, and the objective is to find the most likely one $\hat{\mathbf{e}}$. Therefore, the translation problem can be formalized as formula 2.1 [5].

$$\hat{\mathbf{e}} = \underset{\mathbf{e}}{\operatorname{argmax}} \ \Pr(\mathbf{e} \mid \mathbf{f}) \tag{2.1}$$

Following this expression, a huge data is required to do very good probability estimates of $\Pr(\mathbf{e} \mid \mathbf{f})$, which is almost impossible. Therefore, we break it apart by using Bayes' Rule:

$$\hat{\mathbf{e}} = \underset{\mathbf{e}}{\operatorname{argmax}} \ \frac{\Pr(\mathbf{f} \mid \mathbf{e}) \cdot \Pr(\mathbf{e})}{\Pr(\mathbf{f})} = \underset{\mathbf{e}}{\operatorname{argmax}} \ \Pr(\mathbf{f} \mid \mathbf{e}) \cdot \Pr(\mathbf{e}) \tag{2.2}$$

Equation 2.2 is the fundamental equation of SMT, and the denominator $\Pr(\mathbf{f})$ can be ignored since it is constant for any input source sentence. Corresponding to Figure 2.2, $\Pr(\mathbf{e})$ is the *language model*, and $\Pr(\mathbf{f} \mid \mathbf{e})$ is the *translation model*. Now, the translation problem becomes estimating two probabilities, and devising an optimal search for a target language sentence that maximizes the product of these two models. Note that, although the translation desire is to obtain $\mathbf{e}$ given $\mathbf{f}$, the actual translation model is built reversely. The reason is to use two models to disambiguate $\mathbf{e}$ and counterbalance their errors [30]. In this work we focus on aspects of the translation model that are related to finding the

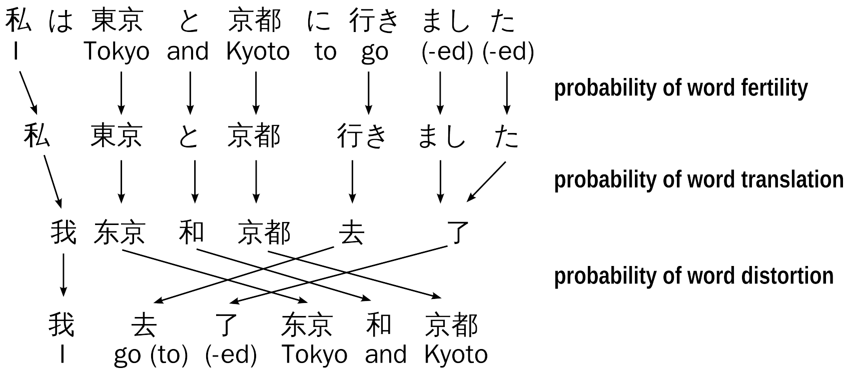


FIGURE 2.3: A simple generation process example from a Japanese sentence (target language) to a Chinese sentence (source language).

appropriate word-to-word correspondence between words of the source language (Chinese) to words of the target language (Japanese).

In order to estimate $\Pr(\mathbf{f} \mid \mathbf{e})$, a parallel corpus is used to train the translation model. In fact, since most of sentences in a corpus appear only once or few times regardless of the corpus size, it is impractical to learn $\Pr(\mathbf{f} \mid \mathbf{e})$ from full sentences. Thus, to computing the conditional probability $\Pr(\mathbf{f} \mid \mathbf{e})$, the strategy is to break the rewriting process from target sentence to source sentence into small steps, and then the probabilities for each steps can be learned. Figure 2.3 illustrates such a decomposed process.[1] To simplify and preserve the correspondences between words in source and target sentences, a hidden variable $\mathbf{a}$ is imported, and the likelihood of the translation $(\mathbf{f}, \mathbf{e})$ in Equation 2.2 can be written in terms of the conditional probability $\Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e})$:

$$\Pr(\mathbf{f} \mid \mathbf{e}) = \sum_{\mathbf{a}} \Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) \tag{2.3}$$

This hidden variable $\mathbf{a}$ stands for *word alignment*, which represent that given a bilingual sentence pair $< \mathbf{f}, \mathbf{e} >$, where $\mathbf{f} = f_1^L$ has l words and $\mathbf{e} = e_1^M$ has m words, then $\mathbf{a} = a_1^M$ is a series of m values, and each between 0 and L. Therefore, $a_j = i$ if f_j aligns with e_i, and $a_j = 0$ means that there is no e_i that is connected to f_j.

[1] Note that this is different from IBM Model 3 since IBM Model 3 also models NULL insertion.

2.1.3 The Word Alignment Problem

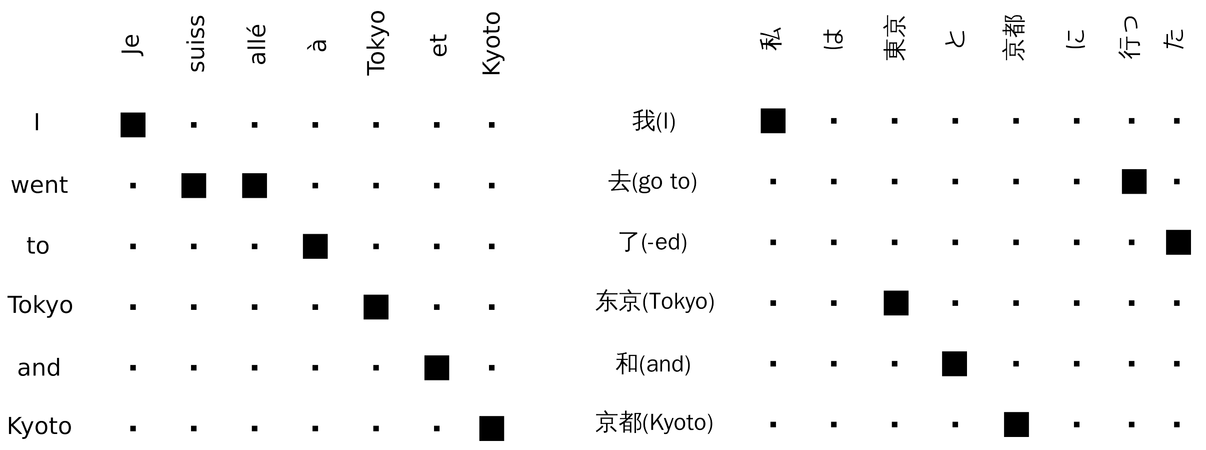
As a combinatorial task, all alignments are deemed possible in a given bilingual sentence pair. Therefore, in order to distinguish the alignment quality, alignment probability is assigned to each particular alignment given a certain sentence pair. Mathematically, the alignment conditional probability is:

$$\Pr(\mathbf{a} \mid \mathbf{e}, \mathbf{f}) = \frac{\Pr(\mathbf{f}, \mathbf{a}, \mathbf{e})}{\Pr(\mathbf{f}, \mathbf{e})} = \frac{\Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e})}{\Pr(\mathbf{f} \mid \mathbf{e})} \tag{2.4}$$

Both Equation 2.3 and 2.4 show that the key is to compute the joint probability of a particular alignment and a source word sequence given an target sentence, namely $\Pr(\mathbf{f}, \mathbf{a} \mid \mathbf{e})$, and it is the product of a group of probabilities, such as the ones shown in Figure 2.3. Brown et al. [5] used the *Estimation-Maximization* (EM) algorithm [33, 34] to optimize parameter values and converge to a local maximum of the likelihood of a particular set of translations which is the so-called *training data*. They proposed several alignment models, namely IBM models, and currently IBM Model 4 is widely used as the final word alignments output. However, as the authors in [5] noted, due to a combinatorial problem, there is not a known method to estimate the probabilities of every possible alignment configuration. For this reason, only a subset of the possible alignment configurations is explored. This subset consists in an initial guess of the best alignment and its neighboring alignment hypotheses, obtained by performing small variations in the configuration of the alignment.

The IBM models are originally designed for translation between English and French. This pair of languages, although have some small differences in word order (such as the relative position of adjectives and nouns)[2], do not contain significant structural differences. When considering Chinese and Japanese as a language pair, however, we observe important structural differences that lead to long range word order differences. For this reason, the suboptimal solution of using the subset of neighboring alignments might not be appropriate in the task of recognizing word correspondences between Chinese and Japanese sentences. Alignment matrices in Figure 2.4 illustrates differences in word alignment

[2]adjective + noun in English, noun + adjective in French



(A) Alignment in English and French

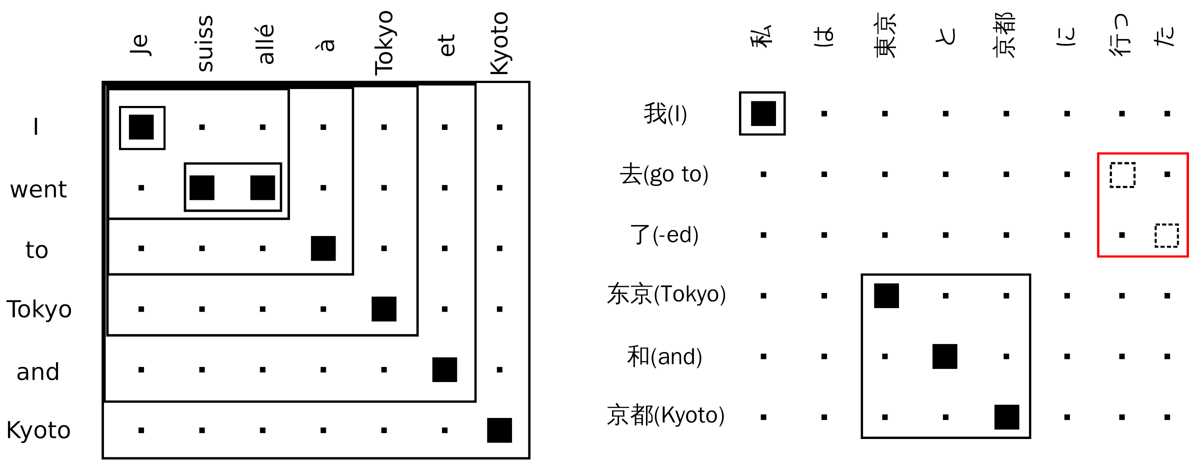
(B) Alignment in Chinese and Japanese

Figure 2.4: Word alignments between bilingual sentences. Figure 2.4a shows a monotonic word alignment between words in an English sentence and words in a French sentence. Figure 2.4b shows a non-monotonic word alignment between words in Chinese and Japanese sentences. We can observe long-distance word alignments that might be difficult to capture with IBM Models.

between French and English, and between Chinese and Japanese. In the French and English example, there is a monotonic word alignment between words of their respective sentences due to similarities in their sentence structure. In the case of Chinese and Japanese, there are gaps in the matrix of alignments corresponding to Chinese words that align to Japanese words, because they are in very different positions in their respective sentences. An example of such a gap is the alignment between the Chinese words "去(go to) 了(-ed)" and the Japanese words "行っ(go) た(-ed)".[3]

In the training stage, these gaps in the matrix of alignments caused by long distance order differences may lead to IBM Model 4 to miss such word correspondences. Furthermore, the wrong alignment will causes problem during phrase extraction. In the decoding stage, these severe order differences may also fail in finding appropriate translations to source words whose translation would be mapped to a different position in the translated sentence.

[3]In this thesis, we represent Chinese and Japanese characters with their English translation in bracket, e.g., 我(I) for Chinese, and 私(I) for Japanese.



(A) Phrases in English and French

(B) Phrases in Chinese and Japanese

FIGURE 2.5: Phrase extraction given word alignments. Figure 2.5a shows possible bilingual phrases extracted from a monotonic word alignment in English and French. Figure 2.5b shows some extracted phrases from a non-monotonic word alignment between Chinese and Japanese. Legitimated bilingual phrases may not be extracted if word alignments are not accurate.

2.1.4 Phrase Extraction

Current state-of-the-art SMT systems find the best translation of $\mathbf{f}$ by modeling the posterior probability $\Pr(\mathbf{e} \mid \mathbf{f})$ directly, and maximum entropy [35] for statistical modeling is a well-founded framework. This method is the so-called log-linear models [36–38] where the decision rule is given by

$$\begin{aligned} \hat{\mathbf{e}} &= \operatorname*{argmax}_{\mathbf{e}} \frac{\exp \sum_{n=1}^{N} \lambda_n h_n(\mathbf{f}, \mathbf{e})}{\sum_{\mathbf{e}'} \exp \sum_{n=1}^{N} \lambda_n h_n(\mathbf{f}, \mathbf{e}')} \\ &= \operatorname*{argmax}_{\mathbf{e}} \sum_{n=1}^{N} \lambda_n h_n(\mathbf{f}, \mathbf{e}), \end{aligned} \tag{2.5}$$

In Equation 2.5, $h_n(\mathbf{f}, \mathbf{e})$, $n = [1, \ldots, N]$, is a set of N feature functions for the translation of $\mathbf{f}$ into $\mathbf{e}$. λ_n are the model parameters for each feature function.

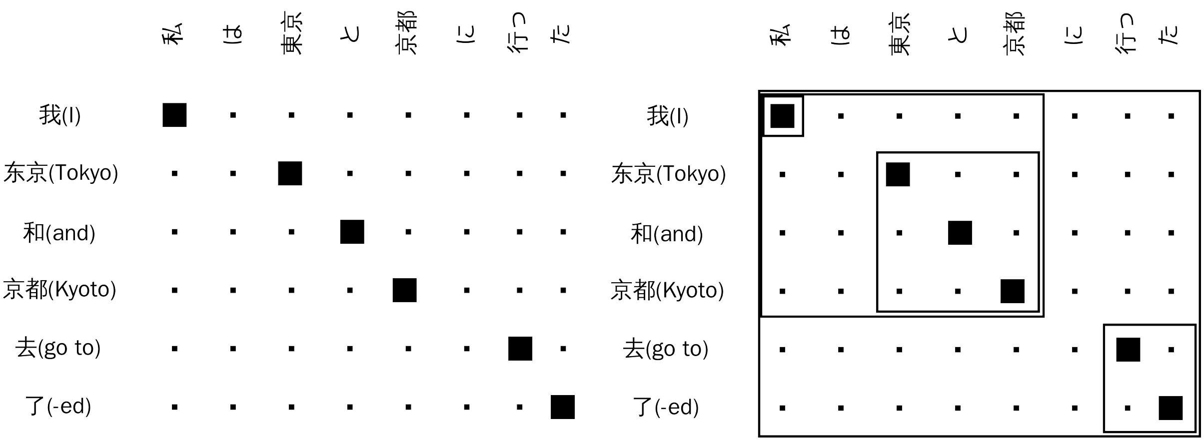
Among these N models, there are translation models, which are nothing else than large tables where each entry contains a phrase in the source language, a phrase in the target language, and the frequency that such a phrase was found in the bilingual training corpus. The constructions of these phrase tables heavily relies on accurate word-to-word alignments, as the ones shown in Figure 2.4. However, due to long distance word order differences, IBM models may fail at recognizing the correct word alignment.

Figure 2.5 depicts a simple example of phrase extraction in English-French and Chinese-Japanese, given the word alignments in bilingual sentences. In Figure 2.5a, bilingual phrase pairs such as "I, Je", and "I went to, Je suiss allé" would be extracted and added to the phrase table, and its direct and inverted frequency would be computed over all occurrences of source and target phrases. In Chinese-Japanese, the bilingual phrase "我(I), 私(I)" would be successfully extracted. However, if IBM models fail to recognize the alignment between the phrase "去(go to) 了(-ed)" and "行っ(go) た(-ed)", such a phrase pair would not be successfully extracted (shown in red on Figure 2.5b). In the best case, the Chinese words "去(go to) 了(-ed)" would not be aligned to any other Japanese word, resulting in a lack of phrase coverage in the models. In the worst case, the Chinese words "去(go to) 了(-ed)" would be aligned to non-corresponding Japanese words, resulting in the extraction of wrong bilingual phrases, leading to a lack of precision of the machine translation system.

As we have observed, languages with different sentence structures pose additional challenges in the word alignment problem, which may result in phrase tables with lack of coverage or diminished precision. To alleviate this problem, *pre-reordering* is a popular technique that aims to pre-process the source language (Chinese) to produce Chinese sentences with a word order that resembles that of Japanese sentences.

Figure 2.6 illustrates such pre-processing. Words in the Chinese sentence are re-arranged to resemble the word order of the Japanese sentence. Thus, the IBM models are likely to find the correct word alignment since the true and hidden alignment is monotonic. Then, the phrase extraction also would have more chances to extract legitimated bilingual phrases such as "我(I), 私(I)", "我(I) 东京(Tokyo) 和(and) 京都(Kyoto), 私(I) は 東京(Tokyo) と(and) 京都(Kyoto)", or "去(go to) 了(-ed), 行っ(go) た(-ed)".

Pre-reordering may play an important role in translation models, since it allows to easily introduce hand-crafted or automatically extracted reordering rules in the statistical machine translation system.. In this thesis, we show how to introduce linguistic intuitions



(A) Alignment in Chinese and Japanese (B) Phrases in Chinese and Japanese

FIGURE 2.6: Alignment and phrase extraction after pre-reordering. Figure 2.6a shows a monotonic word alignment between Chinese and Japanese words after pre-reordering the Chinese sentence. Figure 2.6b shows some extracted phrases from this new monotonic word alignment. The phrase pair "去(go to) 了(-ed), 行っ(go) た(-ed)" is more likely to be successfully extracted.

on ordering differences in translation models that participate in the log-linear combination of Equation 2.5. Our linguistic intuitions will be formalized as reordering rules, that will be used to reorder Chinese sentences in the training corpus. During the decoding stage, unseen Chinese sentences will also be similarly pre-reordered following the same reordering rules, and translated into the Japanese sentences normally using the phrases that we have previously extracted.

2.2 Parsing

In theoretical linguistics, parsing can be distinguished into several types in terms of the formal grammar, e.g., phrase structure grammars, dependency grammars. Linguistically motivated pre-reordering models obtain syntactic information by using source language parsers. There are mainly two types of parsers that have been used to extract sentence structure and guide reordering. The first type corresponds to parsers that extract phrase structures (i.e. Head-driven phrase structure grammar parsers). These parsers infer a rich annotation of the sentence in terms of syntactic or semantic structure. Other reordering strategies use a different type of parsers, namely dependency parsers. These parsers extract dependency information among words in the sentence, often consisting in the dependency relation between two words and the type of relation (dependency label).

In our first contribution, we adapt and refine for Chinese an existing pre-reordering method Head Finalization (HF) [1] which was using *Enju*[4] [39], an HPSG based deep parser for English. We follow their observation and accordingly use the Chinese HPSG based parser *Chinese Enju* [40] for Chinese syntactic parsing.

An XML format output example of *Chinese Enju* for the Chinese sentence "我(I) 去(go to) 了(-ed) 东京(Tokyo) 和(and) 京都(Kyoto)." is given in Figure 2.7. Label "`<cons`" represents non-terminal node while label "`<tok`" represents terminal node. Each node is identified by an unique "`id`" and has several attributes in which the attribute "`head`" indicates its syntactic head. As an example, the first line in Figure 2.7 defines a non-terminal node whose id is "c1" and whose syntactic head is node "t0". Based on the binary tree structure and head information produced by the parser, a simple swapping tree operation can reorder a head-initial language like Chinese to follow a head-final word order.

In our second contribution, we use an unlabeled dependency parser for Chinese, *Corbit*[5] [41] which is based on dependency grammar. Unlike phrase structure grammars, dependency grammar has a flatter structure since it is determined by the relation between a word and its dependents. Figure 2.8 gives an example of unlabeled dependency parse tree.

2.3 Resource

In order to evaluate the performance of our pre-reordering methods for Chinese to Japanese machine translation not only in a single domain but also in multiple domains, we collected corpora from two domains: news and patent, and we used two corpora for each domain. As for the news domain, we obtained an in-house Chinese-Japanese parallel corpus of news articles that we call *News*, and used it as a training set (Training 1).

[4] http://www.nactem.ac.uk/enju
[5] http://triplet.cc/software/corbit

```
<cons id="c1" cat="N" head="t0">
  <tok id="t0" cat="N" pos="PN">wo3 我 (I)</tok>
</cons>
<cons id="c2" cat="V" head="c3" schema="head_mod">
  <cons id="c3" cat="V" head="c4" schema="head_comp">
    <cons id="c4" cat="V" head="c5" schema="head_marker">
      <cons id="c5" cat="V" head="t1">
        <tok id="t1" cat="V" pos="VV" arg1="c1" arg2="c7">qu4 去 (go to)</tok>
      </cons>
      <cons id="c6" cat="MARK" head="t2">
        <tok id="t2" cat="MARK" pos="AS" arg1="c5">le0 了 (-ed)</tok>
      </cons>
    </cons>
    <cons id="c7" cat="N" head="c8" schema="coord_left">
      <cons id="c8" cat="N" head="t3">
        <tok id="t3" cat="N" pos="NR">dong1jing1 东京 (Tokyo)</tok>
      </cons>
      <cons id="c9" cat="COOD" head="c10" schema="coord_right">
        <cons id="c10" cat="CONJ" head="t4">
          <tok id="t4" cat="CONJ" pos="CC" arg1="c8" arg2="c11">he2 和 (and)</tok>
        </cons>
        <cons id="c11" cat="N" head="t5">
          <tok id="t5" cat="N" pos="NR">jing1du1 京都 (Kyoto)</tok>
        </cons>
      </cons>
    </cons>
  </cons>
  <cons id="c12" cat="PU" head="t6">
    <tok id="t6" cat="PU" pos="PU">。</tok>
  </cons>
</cons>
```

FIGURE 2.7: An XML format output of *Chinese Enju* for a Chinese sentence. For clarity, we only draw information related to the phrase structure and the heads.

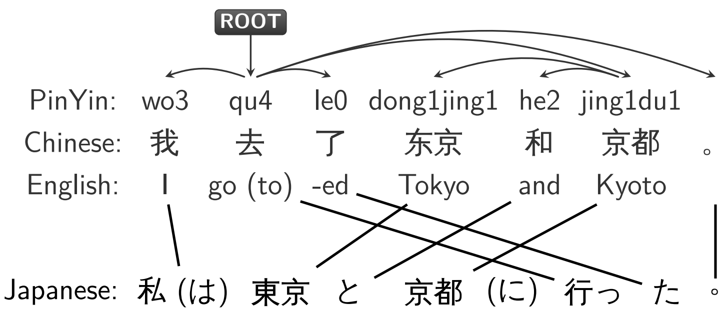


FIGURE 2.8: An example of unlabeled dependency parse tree graph of a Chinese sentence with word aligned to its Japanese counterpart. Arrows are pointing from heads to their dependents.

Chinese sentences in this corpus are extracted from *Xinhua* news in the period of August 1, 2003 − July 31, 2005, and then human translators rendered them into Japanese. We augmented this corpus with another corpus which is from the 7th China Workshop on Machine Translation (CWMT2011)[6] [42], to use it as an extended training set (Training 2). For the patent domain, the corpora were extracted from patent applications filed from 2007 to 2010. The document alignment was based on the priority claims and the sentence alignment was done using Champollion Tool Kit (CTK)[7] [43]. CTK requires a bilingual lexicon, and uses it to find anchors between a candidate source sentence and a possible target sentence. These anchors do not contribute with the same weight to the decision of whether two sentences are translations of each other. Instead, they give a larger weight to less frequent tokens (such as numbers or other symbols), and a lower weight to very frequent tokens, such as punctuation marks. Then, CTK uses dynamic programming to find the minimum sequence of edits to convert the source sentence into the target sentence, giving some empirical costs to deletion and insertion operations. We extract sentence pair whose alignment score is higher than 0.95 to build up Training 1 and alignment score is higher than 0.9 to build up Training 2, which means that training 2 includes all sentences in Training 1 but is larger. Finally, for every domain we obtained a disjoint set of sentences for development and test. Statistics on these corpora can be found in Table 2.1. Out of vocabulary words are computed with respect Training 1 and Training 2, respectively.

[6]http://mt.xmu.edu.cn/cwmt2011/document/papers/e00.pdf
[7]http://champollion.sourceforge.net/

TABLE 2.1: Statistical Characteristics of Corpora

		News		Patent	
		Chinese	Japanese	Chinese	Japanese
Training 1	Sentences	342,050		2,559,581	
	Running words	7,414,749	9,361,867	64,028,414	78,624,671
	Avg. sent. len.	21.68	27.37	25.02	30.72
	Vocabulary	145,133	73,909	351,345	91,778
Training 2	Sentences	621,610		4,894,415	
	Running words	9,822,535	12,499,112	132,206,053	164,452,302
	Avg. sent. len.	15.80	20.11	27.01	33.60
	Vocabulary	214,085	98,333	526,545	124,512
Development	Sentences	1,000		1,144	
	Running words	46,042	56,748	31.57	39.71
	Avg. sent. len.	46.04	56.75	36,114	46,570
	Out of Vocab.	301 and 262	67 and 57	112 and 73	26 and 20
Test	Sentences	2,000		1,144	
	Running words	51,534	65,721	37,145	47,750
	Avg. sent. len.	25.77	32.86	32.47	40.74
	Out of Vocab.	594 and 546	310 and 278	100 and 53	23 and 9

Chapter 3

Related Work

Using statistical techniques to solve the problem of machine translate on natural languages can date back to 1949 when Warren Weaver proposed [15]. Later on, while significant improvement had been achieved by moving from word-based model to phrase-based model based SMT system (PSMT), as one of the most disturbing issues that causes the decrease of translation quality in long distance language pairs, word order difference catches more and more attention in this community. Reordering becomes a popular strategy. From language aspect, there are mainly two families: namely, language independent reordering models and language dependent reordering models [44]. In this chapter, we will briefly introduce previous works on both and also related works on analyzing the relation between parsing and reordering.

3.1 Language Independent Reordering

Over the past a few decades, many researchers in the machine translation community have been working on improving word alignments. Back to the time of word-based models for SMT in 1990's, early approaches [5, 30, 45] built SMT systems precisely based on the statistical models that we introduced in Section 2.1.2. As for the alignment model (equation 2.3), the map of word position from the source language to the target language is treated as a hidden variable. These early systems, instead of modeling word reordering, were using widely word-to-word distance probabilities to overcome non-monotonic word order problems, and this relative distance reordering model was designed to heavily penalize the move of a word over a long distance in a sentence. Moreover, since the reordering was largely induced by the language model, an n-gram language model limited the reordering window to n words. In early days, although trigram was considered as sufficient as building a language model, it is noticeably that three words are inadequate to examine a sentence grammatically [14].

Later in phrase-based modeling [7, 8, 38], which is generally considered as a major breakthrough in SMT, word order differences between source and target language gain more and more attention in machine translation pipeline. Since searching over all reordering possibilities during translation would be NP-hard problem [46], state-of-the-art phrase-based SMT systems are based on beam-search algorithm [47, 48]. Generally speaking, unsupervised word alignment methods [5, 7, 8, 49] performed reasonably well for language pairs that are structurally similar to each other. However, since local differences in word order still exist, which make word alignments non-monotonic, efforts have been on addressing such issues, and lexicalized reordering model [50, 51] was proposed based on the phrase table (as an example in Figure 2.6b). To avoid data sparseness, the widely used lexicalized reordering model [52] only considers three types of orientation, i.e., monotone, swap, and discontinuous.

As one of the flat reordering models, the work in [53] focused on local phrase reorderings, and implemented the models with weighted finite state transducers. The difference with other works is that their parameters were estimated during the phrase alignments and used a EM-style method. There have been also works on using N-best phrase alignment and phrase clustering to estimate the distortion probabilities, and aiming for long-distance reordering [54, 55]. Authors in [56] proposed a distortion model which learns the parameters directly from word alignments, and the probability distributions for each of three distortions in the model were conditioned on the source words. To better deal with the data sparseness, Zens and Ney [57] introduced more features, i.e., words, word classes, local context, into their discriminative reordering model, and used the framework of maximum entropy to handle all different features. Differently, Xiong et al. [58] used the idea of maximum entropy to build a classification model since they only considered types of reorderings, *straight* and *inverted*. A recent work on lexicalized reordering model was presented by [59] in which the authors trained a syntactic analogue of a lexicalized reordering model by using multiword syntactic labels from Combinatory categorial grammar parse charts for Urdu-to-English translation.

3.2 Language Dependent Reordering

Linguistically motivated pre-reordering method usually involves a parser/parsers via which to obtain syntactic information of either source/target language or both, and this method has been proved to be an efficient auxiliary technique for a traditional phrase-based SMT system to improve the translation quality by many researches [60–65], especially when source and target languages are structurally very different, such as German-English [61], English-Arabic [66], English-Hindi [67], Japanese-English [1, 68], English to multiple SOV/VSO languages (i.e., Korean, Japanese, Hindi, Urdu, Welsh and Turkish) or noun-modifier issues (i.e., Russian and Czech) [63, 64] and so on. As for Chinese-to-Japanese translation, there are limited previous works but using pivot language [69, 70]. To the best of our knowledge, this is the first work of using reordering method for Chinese-to-Japanese SMT.

From the generated parse trees, researchers adopt two main strategies to extract reordering rules. One is to create handcrafted reordering rules based on linguistic analysis [61–63, 68], whereas another one is to learn reordering rules from the data [60, 64, 65, 71, 72]. Xia and McCord [60] presented a method to automatically learn rewriting patterns from the combination of aligned phrases and their parse tree pairs. Li et al. [71] used tree operations to generate an n-best list of reordered candidates from which to produce the optimal translation. In [64], reorderings were operated on shallow constituent trees which were converted from dependency parse trees, and reordering rules were extracted automatically from aligned bitext.

For our first method, we are centered in the design of manual rules inspired by the Head Finalization (HF) pre-reordering method described in [1]. HF pre-reordering method is one of the simplest methods that significantly improves word alignments and leads to a better translation quality. The implementation of HF method for English-to-Japanese translation appeared to work well. A reasonable explanation for this is the close match of the syntactic concept *head* in such language pair. But for Chinese-to-Japanese, differences in the definition of *head* lead to unexpected reordering problems while implementing HF. As we believe that such differences are also likely to be observed in other language pairs. A more detailed description of HF will be in Section 3.4.

Even though our refined HF method for Chinese (HFC) has produced gains in reordering quality, it is impractical to add enormous handcrafted rules to solve infinite reordering issues. We hence propose another new pre-reordering framework for Chinese based on unlabeled dependency parsing (DPC). One similar work as ours was introduced in [63], which was using an English dependency parser to formulate handcrafted reordering rules in the form of triplet that is composed of dependency labels, part-of-speech (POS) tags and weights. The rules were operated recursively in a sentence while reordering. This design, however, limits the extensibility of their method. Our approach follows the idea of using dependency tree structures and POS tags, but we discard the information on dependency labels since we did not find them informative to guide our reordering strategies in our preliminary experiments, partly due to Chinese showing less dependencies and a larger label variability [73].

Another direction of pre-reordering is to develop reordering rules without using a parser [74–78]. For instance, in [74], reordering source language was treated as a translation task in which statistical word classes were used; but in [75], reordering rules were learned from POS tags instead of parse trees; authors in [76] and [77] proposed methods of using binary classification; and Neubig et al. [78] presented a traditional context-free-grammar models based method for learning a discriminative parser to improve reordering accuracy.

Although the majority of efforts were dedicated to pre-reordering, other authors [79, 80] examined the possibility of post-reordering for a Japanese-to-English translation task. Post-ordering could be seen as related to post-editing technologies but essentially different. The authors first translated Japanese to Japanese-ordered English, and then reordered this Japanese-ordered English to normal English with an existing reordering method. Comparing with pre-reordering, post-reordering needs two types reordering rules, which are 1) reordering English to Japanese-ordered English for training data and 2) reordering Japanese-ordered English to common English.

3.3 Error Analysis

Besides reordering methods, one of our main contributions is on observing the relationship between parsing errors and reordering errors, which is likewise the first work as far as we know. Although there are studies on analyzing parsing errors or translation errors, there is not any work on observing the relationship between parsing errors and reorderings.

One most relevant work to our analysis work is observing the impact of parsing accuracy on a SMT system introduced in [81]. They showed the general idea that syntax-based SMT models are sensitive to syntactic analysis. However, they did not further analyze concrete parsing error types that affect task accuracy.

Green [82] explored the effects of noun phrase bracketing in dependency parsing in English, and further on English to Czech machine translation. But the work focused on using noun phrase structure to improve a machine translation framework. In the work of [83], they proposed a training method to improve a parser's performance by using reordering quality to examine the parse quality. But they did not study the relationship between reordering quality and parse quality.

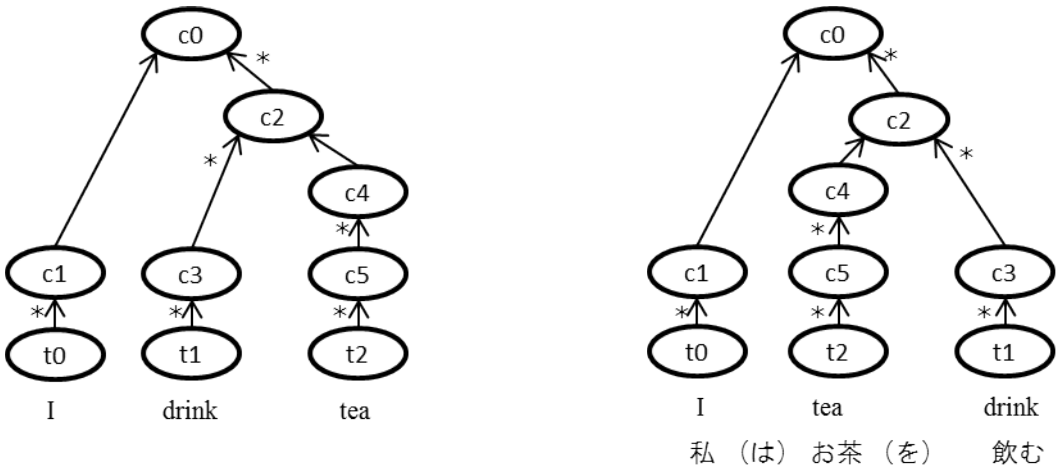
There are more works on parsing error analysis. For instance, [84] defined several types of parsing error patterns on predicate argument relation and tested them with a Head-driven phrase structure grammar (HPSG). [85] explored parsing errors for data-driven dependency parsing by comparing a graph-based parser with a transition-based parser, which are representing two dominant parsing models. At the same time, [86] provided a comparison analysis on differences in annotation guidelines among treebanks which were suspected to be responsible for dependency parsing errors in domain adaptation tasks. Unlike analyzing parsing errors, authors in [40] focused on the difficulties in Chinese deep parsing by comparing the linguistic properties between Chinese and English.

On the other hand, [87] proposed an automatic error analysis method of machine translation output, by compiling a set of metric variants. However, they did not provide insight on what SMT component caused low translation performance.

3.4 Head Finalization (HF)

The structure of languages can be characterized by phrase structures. The head of a phrase is the word that determines the syntactic category of the phrase, and its modifiers (also called dependents) are the rest of the words within the phrase. In English, the head of a phrase can be usually found before its modifiers. For that reason, English is called a head-initial language [88]. Japanese, on the other hand, is head-final language [89], since the head of a phrase always appears after its modifiers.

In certain applications, as in the case of machine translation, word reordering can be a promising strategy to ease the task when working with languages with different phrase structures like English and Japanese. Head Finalization (HF) is a successful syntax-based reordering method that was originally designed to reorder sentences from English (head-initial language) to resemble the word order of Japanese (head-final language) [1]. The essence of this rule is to move the syntactic heads to the end of its dependency by swapping child nodes in a phrase structure tree when the head child appears before the dependent child. Therefore, this simple method need to obtain phrase structures and



(A) Original HPSG English parse tree

(B) Reordered English sentence

Figure 3.1: Head finalization for a simple English sentence. Figure 3.1a shows the parse tree of the original sentence; Figure 3.1b is the reordered parse tree, and reordered English along with its Japanese translation. (* indicate the syntactic head).

syntactic head information first, and Isozaki et al. [1] were using a Head-driven phrase structure grammar based deep parser for English.

Figure 3.1 is a simple HF example to illustrate how to swap the nodes with a HPSG parse tree that is generated by *Enju*. Figure 3.1a gives the original English parse tree, while Figure 3.1b shows reordered English with its Japanese translation. In Figure 3.1a, since node `c3` is the left children node and also is the syntactic head of node `c2`, according to HF reordering rule, it has been swapped with its sibling as shown in Figure 3.1b, so that the English sentence becomes head-final and the alignment with its Japanese translation becomes monotonic.

The values of Kendall's τ reported in [1] show that Head Finalization had improved the word alignment between English and Japanese. A more completed score results later reported in [44] from several mainstream evaluation methods indicated that the translation quality had been improved; the scores of *Word Error Rate (WER)* and *Translation Edit Rate (TER)* [90] had especially been greatly reduced.

3.5 Summary

We have described several methods to tackle the problem of word order differences. Generally, they are divided into two groups, language-independent and language-dependent. On one hand, as the most widely-used content independent reordering method, lexicalized reordering models have drawn the attention of many researchers. On the other hand, pre-reordering has proved to be an effective strategy to introduce linguistic or automatic reordering rules into the machine translation pipeline. In this thesis, we describe our analysis of word order differences between Chinese and Japanese, and synthesize our analysis results into linguistically motivated pre-reordering rules. The final objective of our work will be to obtain monotonic word-to-word alignments in bilingual Chinese-Japanese sentences and improve overall machine translation quality.

Chapter 4

Head Finalization for Chinese (HFC)

In statistical machine translation, reordering rules have proved useful in extracting bilingual phrases and in decoding during translation between languages that are structurally different. Linguistically motivated rules have been incorporated into Chinese-to-English [62] and English-to-Japanese [1] translation with significant gains to the statistical translation system. In this chapter, we carry out a linguistic analysis of the Chinese-to-Japanese translation problem and based on a previous work propose one of the first reordering rules for this language pair. Experimental results show substantially improvements.

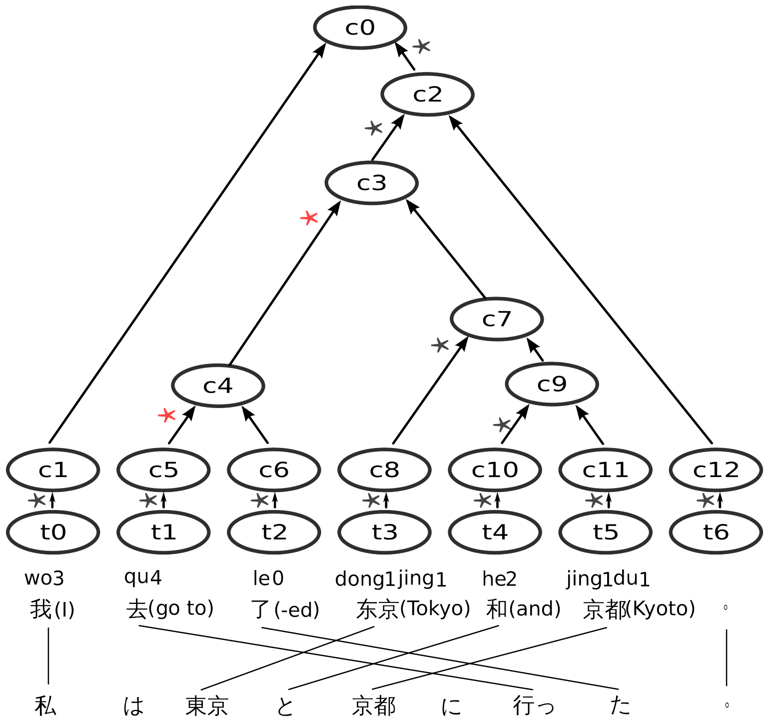
4.1 Preliminary adaptation of HF for Chinese

Since Chinese and English are both known to be head-initial languages[1], in preliminary experiments, we simply adapted HF to Chinese without considering the syntactic distinction between them. Ideally, the reordering rule introduced in Section 3.4 would reorder Chinese sentences following the word order of their Japanese counterparts.

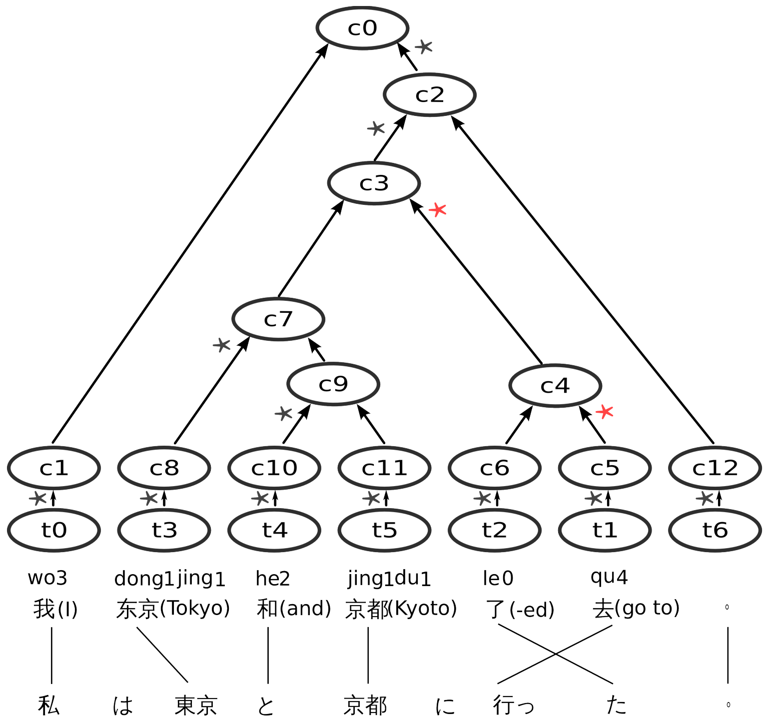
Figure 4.1 shows an example of simple implementing HF to Chinese based on the output of *Chinese Enju* introduced in Figure 2.7 of Section 2.2. Notice that, along with HF, Isozaki et al. [1] have introduced a coordination exception rule and it has been applied to Chinese pre-reordering as well. This exception rule prevents HF from reordering a coordination structure in a sentence and it can be easily implemented by checking a node's attributions of `cat = ''COOD''` or `schema = ''coord-left/right''` as shown in Figure 2.7. Thus, in the example of Figure 4.1a, although nodes of `c8` and `c10` are both the left child nodes and the syntactic heads of `c7` and `c9`, respectively, neither of them were swapped with their siblings as shown in Figure 4.1b. Additionally, although HF does not prevent the swapping operation from crossing punctuation such as commas or quotes, the authors in [1] separated English sentences not only by period, but also by colons and semicolons, which relatively reduced the occurrence of the reordering errors with punctuations. Moreover, unlike the English HPSG parser which does not treat period as part of the parsing tree (See an example tree in Figure 3.1a), the Chinese HPSG parser includes the period in the tree (See an example tree in Figure 4.1a) and the period branch is customarily not the syntactic head. Therefore, we import an punctuation exception rule which terminates all reorderings that involve any punctuation for Chinese reordering. On this account, the node of `c3` did not swap with the node of `c12` in the example. Therefore, only the subtree of verb phrase "去(go to) 了(-ed)" had been moved to the end of the sentence but before the period, which is following a more similar word order as the Japanese translation in Figure 4.1b.

However, a mis-switch took place inside of the verb phrase subtree `c4`. That is, node of `c5` was swapped with node of `c6` due to peculiarities in Chinese syntax and the discrepancies

[1] As Gao [91] summarized, whether Chinese is a head-initial or a head-final language is open for debate. Nevertheless, we take the view that most Chinese sentence structures are head-initial since the written form of Chinese mainly behaves as an head-initial language.



(A) Original HPSG Chinese parse tree



(B) Reordered Chinese sentence

FIGURE 4.1: A simple example of implementing HF to Chinese. Figure 4.1a shows the parse tree of the original Chinese sentence and its English translation. Figure 4.1b shows the reordered Chinese sentence along with its Japanese translation. (* indicate syntactic heads).

in head definition between Chinese and Japanese. For that reason, in the following sections, we will analyze several distinctive cases of the problem in detail and proposes a refinement of the simple adaptation of HF for Chinese pre-reordering with a couple of exception rules.

4.2 Discrepancies in Head Definition

Head Finalization relies on the idea that head-dependent relations are largely consistent among different languages while word orders are different. However, in Chinese, there has been much debate on the definition of head[2], possibly because Chinese has fewer surface syntactic features than other languages like English and Japanese. This causes some discrepancies between the definitions of the head in Chinese and Japanese, which leads to undesirable reordering of Chinese sentences. Specifically, in preliminary experiments we observed unexpected reorderings that are caused by the differences in the head definitions, which we describe below.

4.2.1 Aspect particle

Although Chinese has no syntactic tense marker, three aspect particles following verbs can be used to identify the tense semantically, namely 了(did), 着(doing), and 过(done). Their counterparts in Japanese are た(did), ている(doing), and た(done), respectively. Both the first one and third one can represent the past tense, but the third one is more often used in the past perfect.

The Chinese parser[3] treated aspect particles as dependents of verbs, whereas their Japanese counterparts are identified as the head. The mis-reordering of “去(go to) 了(-ed)” to “了(-ed) 去(go to)” in Figure 4.1 is one of the examples. Since “了(-ed)” is recognized as a dependent of “去(go to)” while its Japanese counterpart is the syntactic head of the verb,

[2]In this thesis, we only consider the syntactic head.

[3]The discussions in this chapter presuppose the syntactic analysis done by *Chinese Enju*, but most of the analysis is consistent with the common explanation for Chinese syntax.

simple implementation of HF leads to the wrong operation. Similarly, 着(doing) and 过(done) are reordered wrongly with the verbs that they are modified.

4.2.2 Adverbial modifier bu4(not)

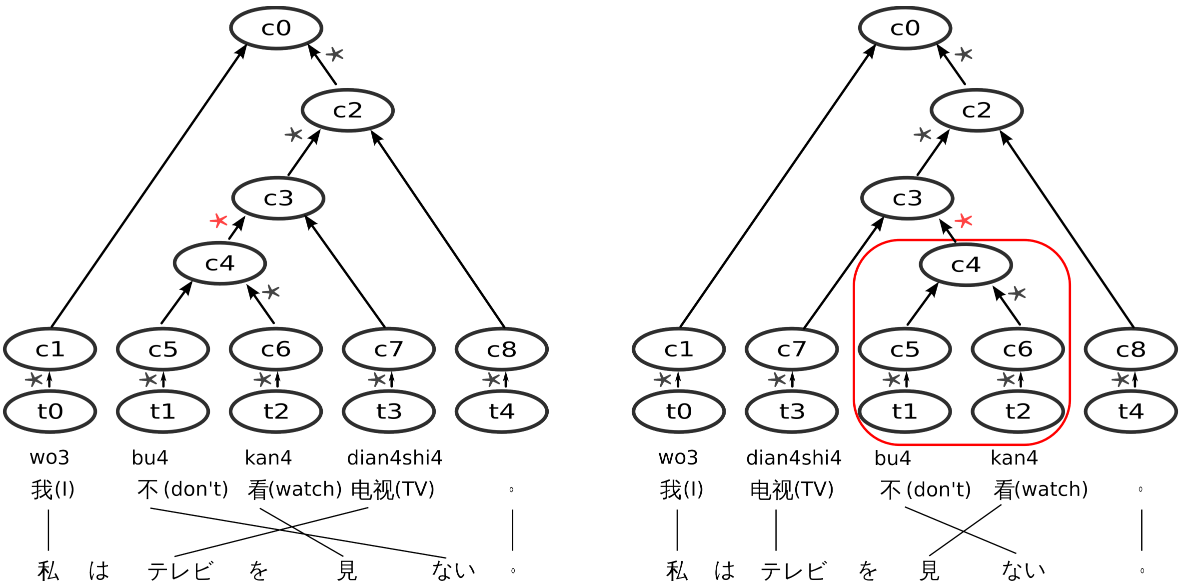
Both in Chinese and Japanese, verb phrase modifiers typically occur in pre-verbal positions, especially when the modifiers are adverbs. Since adverbial modifiers are dependents in both Chinese and Japanese, head finalization works perfectly for them. However, there is an exceptional adverb in Chinese, namely bu4(not) and its Chinese character is 不(not), which functions as a negator and is usually translated into ない in Japanese. As an adverb in Chinese, 不(not) is always a dependent of the verb that it modifies, whereas ない in Japanese is always at the end of the sentence and thus is the syntactic head.

As an illustration, an example is shown in Figure 4.2. In the subtree of `c4`, the verb "看(watch)" is identified as the syntactic head and "不(not)" is its dependent; on the contrary, in the Japanese translation, "ない(not)" as the counterpart of "不(not)" has been identified as the syntactic head of the verb. As a result, the alignment between reordered Chinese sentence and its Japanese translation is not monotonic as shown in Figure 4.2b.

4.2.3 Sentence-final particle

Sentence-final particles often appear at the end of a sentence to express a speaker's attitude: e.g. 吧(right?), 啊(ah) in Chinese, and なぁ, ねえ in Japanese. Although they are in the same position in both Chinese and Japanese, in accordance with the differences of head definition, they are identified as the dependent in Chinese whereas they are the syntactic head in Japanese.

For example in Figure 4.3, since "啊(ah)" was identified as the dependent, it has been reordered to the beginning of the sentence. However, its Japanese translation "ね" is at the end of the sentence and acts as the syntactic head. Likewise, the alignment between



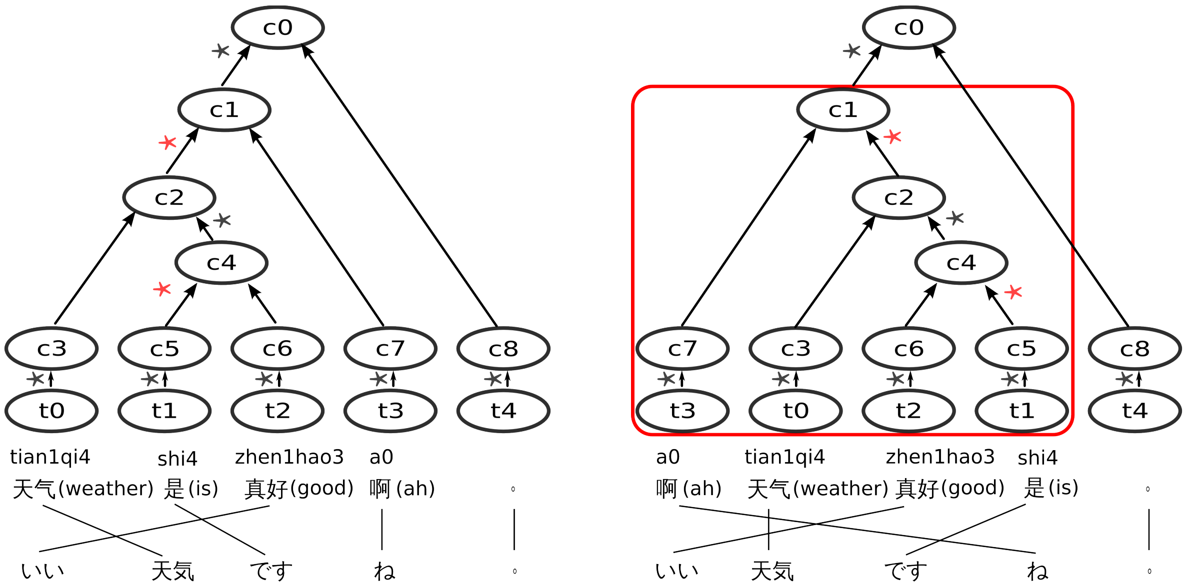
(A) Original HPSG Chinese parse tree (B) Reordered Chinese sentence

FIGURE 4.2: An example of showing the mis-reordering of adverbial modifier 不(not) while implementing HF to Chinese pre-reordering. Figure 4.2a shows the original parse tree and its English translation. Figure 4.2b shows the wrongly reordered Chinese sentence along with its Japanese translation.

reordered Chinese sentence and its Japanese translation is not monotonic as shown in Figure 4.3b.

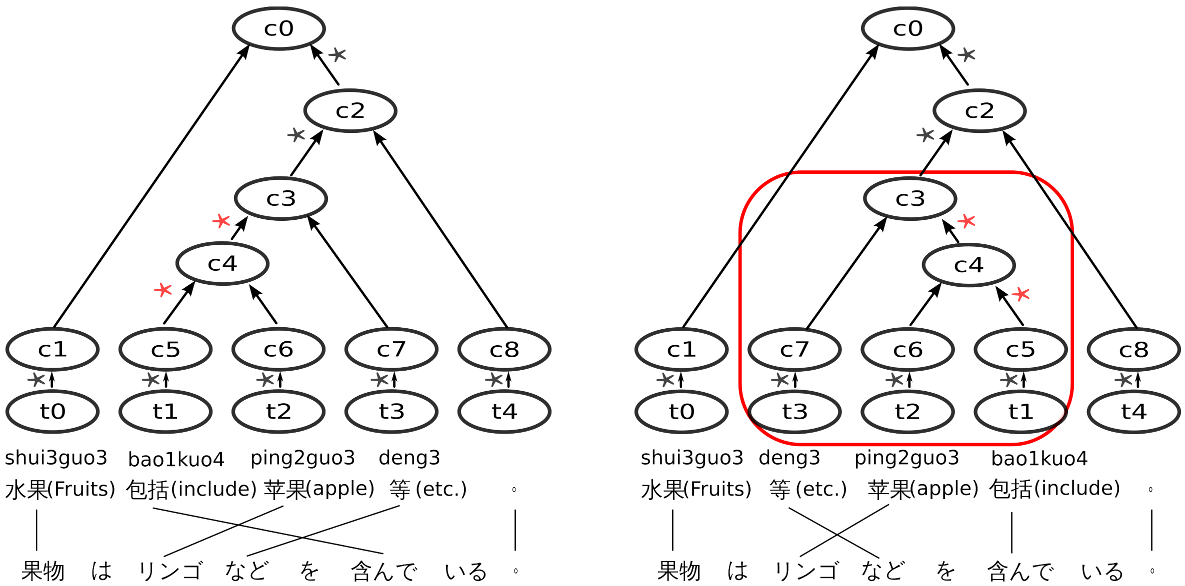
4.2.4 *Et cetera*

In Chinese, there are two expressions for representing the meaning of "and other things" with one Chinese character: 等(etc.) and 等等(etc.), which are both identified as dependent of a noun phrase of verb phrase. In contrast, in Japanese, the translation of *Et cetera* is など which is always the head because it appears as the right-most word in a phrase. For instance, the verb phrase of "包括(include) 苹果(apple) 等(etc.)" in Figure 4.4. Since "等(etc.)" is not the syntactic head in Chinese but is in Japanese, HF produced a wrong reordering for the phrase.



(A) Original HPSG Chinese parse tree (B) Reordered Chinese sentence

FIGURE 4.3: An example of showing the mis-reordering of sentence-final particle while implementing HF to Chinese pre-reordering. Figure 4.3a shows the original parse tree and its English translation. Figure 4.3b shows the wrongly reordered Chinese sentence along with its Japanese translation.



(A) Original HPSG Chinese parse tree (B) Reordered Chinese sentence

FIGURE 4.4: An example of showing the mis-reordering of *Et cetera* while implementing HF to Chinese pre-reordering. Figure 4.4a shows the original parse tree and its English translation. Figure 4.4b shows the wrongly reordered Chinese sentence along with its Japanese translation.

TABLE 4.1: The List of POS tags for Exception Reordering Rules

AS	Aspect particle
SP	Sentence-final particle
ETC	*et cetera* (i.e. 等 and 等等)
IJ	Interjection
PU	Punctuation
CC	Coordinating conjunction

4.2.5 Head finalization for Chinese (HFC)

In the preceding sections, we have discussed syntactic constructions that cause wrong application of Head Finalization (HF) to Chinese sentences. Following the observations, we propose a method to improve the original Head Finalization pre-reordering rule for Chinese to obtain better alignment with Japanese.

The idea is simple: we define a list of Part-of-Speech (POS) tags[4] (See Table 4.1) to control the implementation of HF. In other words, if the POS tag of a leave node in a branch belongs to the predefined POS tag list, whether operates HF on the branch is depended on the exception rules. For example, as for the case of sentence-final particle in Figure 4.3, since the POS tag of "啊(ah)" is SP which is in Table 4.1 and according to the analysis in previous section, HF thus will not be operated on the node of `c1`. That is, branch of `c2` and `c7` will not be swapped.

In Table 4.1, interjection is included as well. Although we did not discuss interjections in detail, it is obviously that interjections should not be reordered, because they are position-independent. Moreover, the rules for PU and CC are basically equivalent to the exception rules proposed by Isozaki et al. [1].

[4]The definitions of POS tags follow the guideline of the Penn Chinese Treebank v3.0 [92].

4.3 Evaluation

4.3.1 Experiment setting

We evaluate the performance of HFC pre-reordering method for Chinese to Japanese machine translation on both corpora from news domain and patent domain that we have described in Section 2.3. Statistics on these corpora can be found in Table 2.1.

Following methods are used to preprocess these parallel corpora to be ready for pre-reordering. We first segmented Japanese and Chinese sentences using *MeCab*[5] [93] and *Stanford Chinese segmenter*[6] [94], respectively. POS tags of Chinese were automatically extracted from the result of *Berkeley parser*[7] [95]. For the method of Head finalization for Chinese (HFC), HPSG parse trees were obtained by using *Chinese Enju* [40].

Previous work on translating from English to Japanese [1] used artificial particle insertions on the source language to align with Japanese particles that do not have a counterpart on the English side. In Chinese to Japanese translation, there are also no Chinese counterparts of Japanese particles. For that reason, we used the same technique as described in [44] to insert particles in Chinese sentences after pre-reordering was carried out for HFC.

The described pre-reordering method (HFC) were applied at a pre-processing stage to increase the chances of monotonic alignments at a later stage. The standard *Moses*[8] [96] baseline was used, where reordered Chinese sentences were paired with their Japanese counterparts and word-to-word alignments were estimated by using *MGIZA++*[9] [9, 97]. Following standard practice, we used the default distance reordering model and "msd-bidirectional-fe" as a lexicalized reordering model [51, 52]. We estimated a 5-gram language model using *The SRI Language Modeling Toolkit (SRILM)*[10] [98] on the target side of the corresponding training corpus. Then, the scaling factors of the log-linear combination of models were estimated using *Minimum Error Rate Training (MERT)* [99].

[5]http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[6]http://nlp.stanford.edu/software/segmenter.shtml
[7]http://nlp.cs.berkeley.edu/Software.shtml
[8]http://www.statmt.org/moses
[9]http://www.kyloo.net/software/doku.php/mgiza:overview
[10]http://www.speech.sri.com/projects/srilm

For comparison reasons, we also included in our evaluation the Head Finalization (HF) system developed in [1] for English to Japanese translation, since English and Chinese are both SVO languages.

Translation quality can be evaluated from different perspectives, and several metrics have been used in the literature for this purpose. *Bilingual Evaluation Understudy (BLEU)* [100] is a widely used metric that computes the geometric average between uni-gram, bi-gram, tri-gram and four-gram precisions, which emphasizes the importance of local fluency. When translating between language pairs with different sentence structure, such as Chinese and Japanese, long distance phrase mis-reorderings are a major source of translation quality drop. *Rank-based Intuitive Bilingual Evaluation Score (RIBES)*[11] [101] is a score based on rank correlations of word orders that was specifically designed to evaluate translation quality between distant language pairs. *Word error rate (WER)*, also known as Levenshtein distance, is an error metric that computes the string distance (at word level) of the translated sentence and the gold sentence. However, *WER* is very sensitive to incorrect word orders, and strongly penalizes machine translated sentences in distant language translations. *Position-independent word error rate (PER)* [102] does not consider the order of words since the position of words in the translated sentence are not taken into account in the error computation. Finally, *Translation Edit Rate (TER)*[12] [90] counts the number of edits that are necessary to convert the machine translated Japanese sentence into the gold Japanese sentence. Possible edit operations are insertions, deletions, substitutions of words and shifts of phrases. *BLEU* and *RIBES* are precision metrics, and higher scores suggest higher performance. Contrarily, *WER*, *PER* and *TER* are error metrics, which means that lower scores are better.

4.3.2 Results

Table 4.3 shows the experimental results on news domain while Table 4.3 lists the The results on patent domain. All the experiments are under the setting of distortion limit as 0. From the experiment results, in News domain, although HF has achieved 0.8

[11] http://www.kecl.ntt.co.jp/icl/lirg/ribes

[12] http://www.cs.umd.edu/ snover/tercom

TABLE 4.2: Evaluation Results of Translation Quality on News Domain. Results are given in terms of BLEU, RIBES, WER, PER, and TER for baseline (BL.), HF, and HFC.

	Training 1					Training 2				
	BLEU	RIBES	WER	PER	TER	BLEU	RIBES	WER	PER	TER
BL.	38.29	84.55	51.93	31.52	48.11	38.35	84.53	52.07	31.39	48.09
HF	39.09	84.77	50.48	30.82	46.44	39.15	84.84	50.59	30.59	46.42
HFC	39.55	84.86	50.00	30.65	45.99	39.54	84.92	50.10	30.52	46.00

TABLE 4.3: Evaluation Results of Translation Quality on Patent Domain. Results are given in terms of BLEU, RIBES, WER, PER, and TER for baseline (BL.), HF, and HFC.

	Training 1					Training 2				
	BLEU	RIBES	WER	PER	TER	BLEU	RIBES	WER	PER	TER
BL.	45.51	84.32	49.03	26.41	41.67	50.83	85.87	45.37	23.71	37.74
HF	44.97	84.19	50.70	27.97	42.34	51.53	85.79	45.59	24.78	37.58
HFC	45.06	84.16	50.40	28.16	42.36	51.77	86.03	45.06	24.36	37.15

improvement on *BLEU* and around 0.3 improvement on *RIBES*, HFC has consistently gained further improvements, around 0.45 on *BLEU* and around 0.1 on *RIBES*. However, on patent domain, neither of HF or HFC performs better than baseline system.

Table 4.4 shows two examples of the reorderings that HF and HFC produced. In the first example (top), the Japanese reference indicates that the particle $_1$等 should appear at the end of its constituent (in purple, underlined with index 1). However, HF incorrectly places that particle at the beginning of the constituent, whereas HFC succeeds in placing that particle at the end, imitating the Japanese word order. In the second example (bottom), we focus our attention to four constituents underlined and indexed as 1, 2, 3 and 4. As we can observe in the Japanese reference, the tokens $_1$で, $_2$2, $_3$た and $_4$た appear after their constituents, but HF incorrectly places the Chinese equivalents to those particles at the beginning of their constituents. In this example, HFC correctly places the particle 了 after the constituents with the indexes of 1, 3 and 4, but incorrectly places the token $_2$2 at the beginning of constituent in the index of 2. Moreover, both HF and HFC fail at moving constituent in the index of 4 to the end of the sentence (as it appears in the Japanese reference) due to the parsing err.

TABLE 4.4: Reordering Examples of HF and HFC from the development set. Constituents are underlined and indexed. Particles of interest appear in red, and their constituents appear in purple.

Japanese-1	L アミノ酸 は 、 $_{1}$ブレビバクテリウム 属 、 コリネバクテリウム 属 、 エシェリヒア 属 等 に 属する 微生物 を 用い た 発酵 法 により 工業 生産 さ れ て いる 。
Original Chinese-1 (English)	L-氨基酸(L-amino acids) 是(is) 通过(by) 使用(use) 属于(belong to) $_{1}$短杆菌属(Brevibacterium) 、 棒状杆菌属(Corynebacterium) 、 埃希氏菌属(Escherichia) 等(etc.) 的(of) 微生物(microorganisms) 的(of) 发酵(fermentation) 方法(method) 来(to) 进行(carry out) 工业(industrial) 生产(production) 的 。
HF-1	L- 氨基酸 $_{1}$等 短杆菌属 、 棒状杆菌属 、 埃希氏菌属 的 微生物 属于 的 发酵 方法 使用 通过 来 的 工业 生产 进行 是 。
HFC-1	L- 氨基酸 $_{1}$短杆菌属 、 棒状杆菌属 、 埃希氏菌属 等 的 微生物 属于 的 发酵 方法 使用 通过 来 的 工业 生产 进行 是 。
Japanese-2	得 られ た クローン の うち 、 相当 数 は T 28 N 変異 を $_{1}$含ん で おり 、 $_{2}$表 2 に 示す よう な T 28 N と 同時に 他 の 変異 も $_{3}$導入 さ れ た 2 重 、 3 重 変異 株 が 多く $_{4}$得 られ た 。
Original Chinese-2 (English)	得到(obtain) 的(-ed) 克隆(clones) 中(among) , 有(there is) 相当(quite) 一 部分(part of) $_{1}$包含(contain) 了(-ed) T28N 突变(mutation) , $_{4}$获得 (receive) 了(-ed) 许多(many) $_{2}$如表(as table) 2 所 示(shown) 那样(as) 的 $_{3}$导入(introduce) 了(-ed) T28N 的 同时(while) 还(also) $_{3}$导入(introduce) 了(-ed) 其他(other) 突变(mutated) 的 2 重(heavy) 、 3 重(heavy) 突变(mutant) 菌株(strain) 。
HF-2	得到 的 克隆 中 , 相当 一 部分 T28N 突变 $_{1}$了 包含 有 , $_{3}$了 许多 $_{2}$2 如表 所 那样 的 $_{3}$导入 示 T28N $_{4}$了 获得 的 同时 还 其他 突变 的 2 重 $_{3}$了 导入 、 3 突变 菌株 重 。
HFC-2	得到 的 克隆 中 , 相当 一 部分 T28N 突变 $_{1}$包含 了 有 , 许多 $_{2}$2 如表 所 那样 的 $_{3}$导入 示 $_{3}$了 T28N $_{4}$获得 了 的 同时 还 其他 突变 的 2 重 $_{3}$导入 了 、 3 突变 菌株 重 。

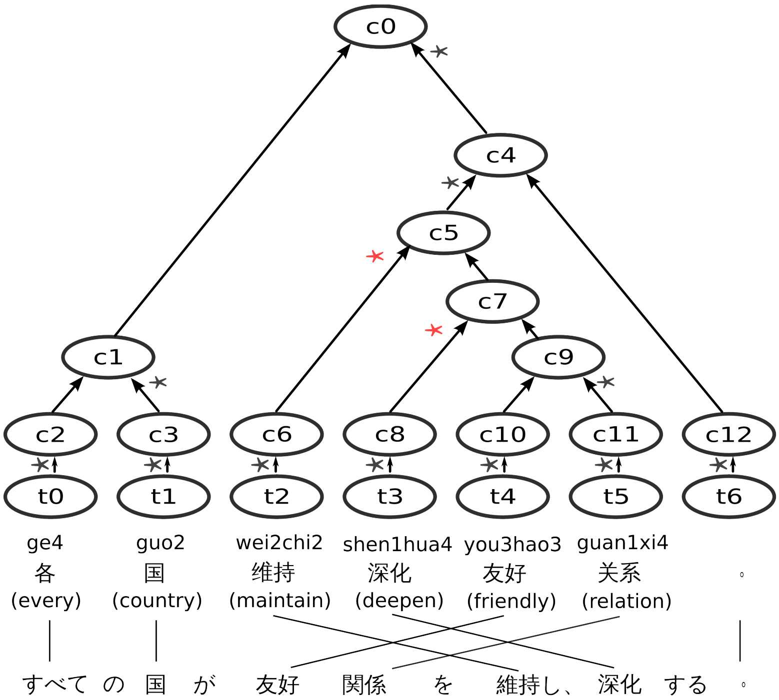
4.4 Error Analysis

In Section 4.2 we have analyzed syntactic differences between Chinese and Japanese that led to the design of an effective refinement of HF for Chinese (HFC). A manual error analysis of the results of HFC showed that some more reordering issues remain. Further improvement on HFC with more exception rules is difficult due to the phrase structure of the HPSG tree and the methodology of HF.

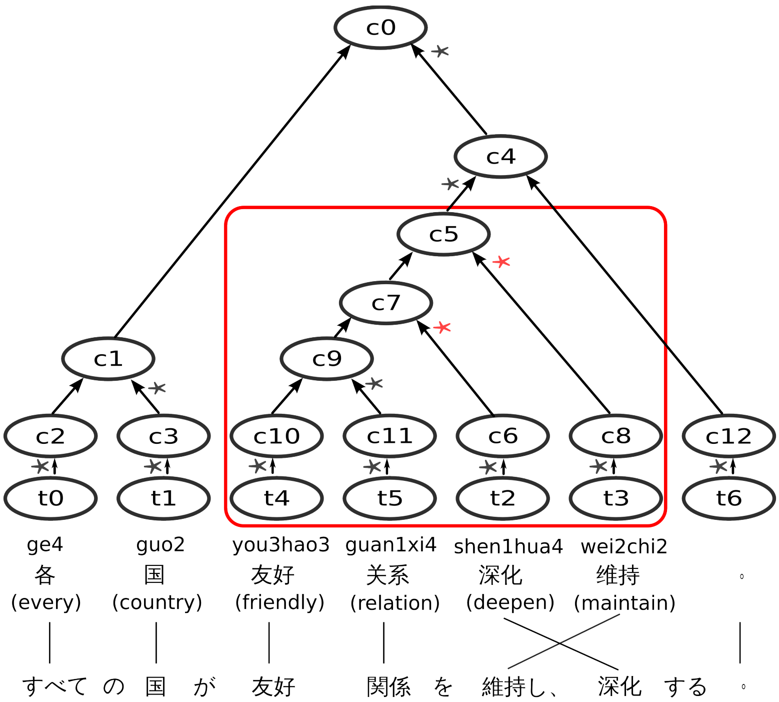
4.4.1 Serial verbs

As discussed by many researchers [103, 104], compared to English and Japanese, Chinese has little inflectional morphology, that is, no inflection to denote tense, case, etc. Therefore, It is crucial to identify the relations among a series of verbs in a Chinese sentence. One of the most common types is called serial verb construction of which the relations among verbs are progressive or parallel. Apparently, there is a largely corresponding construction in Japanese, which indicates that verb sequence as a serial verb construction should be treated as a unity and no reordering should be operated inside of it. Another type of relation among serial verbs is subordinate relationship. That is, a verb phrase acts as an object of the former verb. Therefore, verbs should be reordered separately according to the sentence structure. However, since coordination conjunction words or symbols (i.e., comma) are usually omitted in the first case, it is difficult for a Chinese parser to distinguish these two types of serial verbs, which leads to unexpected reordering results. Examples to illustrate this problem are given in Figure 4.5 and Figure 4.6.

In Figure 4.5, verbs of “维持(maintain)” and “深化(deepen)” compose a serial verb construction in a parallel relation, and they share the same object of “友好(friendly) 关系(relation) ”. Since the word order of this parallel serial verb construction is the same as the word order of its Japanese translation, namely “維持(maintain)” and “深化(deepen)”, there is no reordering requirement between these two verbs. However, the omission of coordination conjunction words or symbols in the construction brings difficulty for Chinese parsers to recognize it. As shown in Figure 4.5a, verb “深化(deepen)” together with the noun phrase of “友好 关系(friendly relation)” is considered as a verb phrase

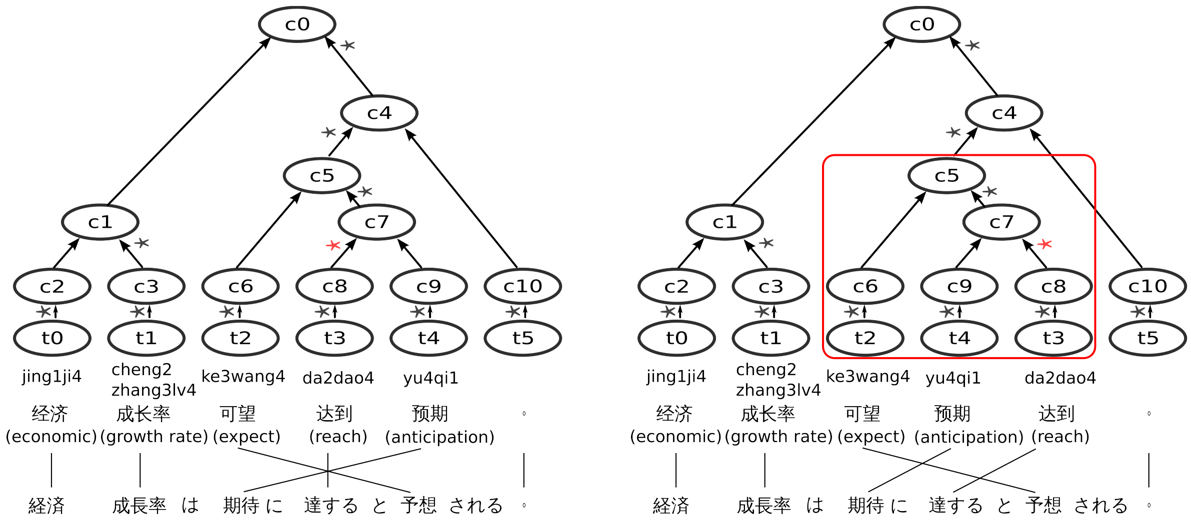


(A) Original HPSG Chinese parse tree



(B) Reordered Chinese sentence

FIGURE 4.5: An example of showing the mis-reordering of serial verb construction while implementing HFC. Figure 4.5a give the original parse tree and its English translation. Figure 4.5b shows the wrongly reordered Chinese sentence along with its Japanese translations.



(A) Original HPSG Chinese parse tree (B) Reordered Chinese sentence

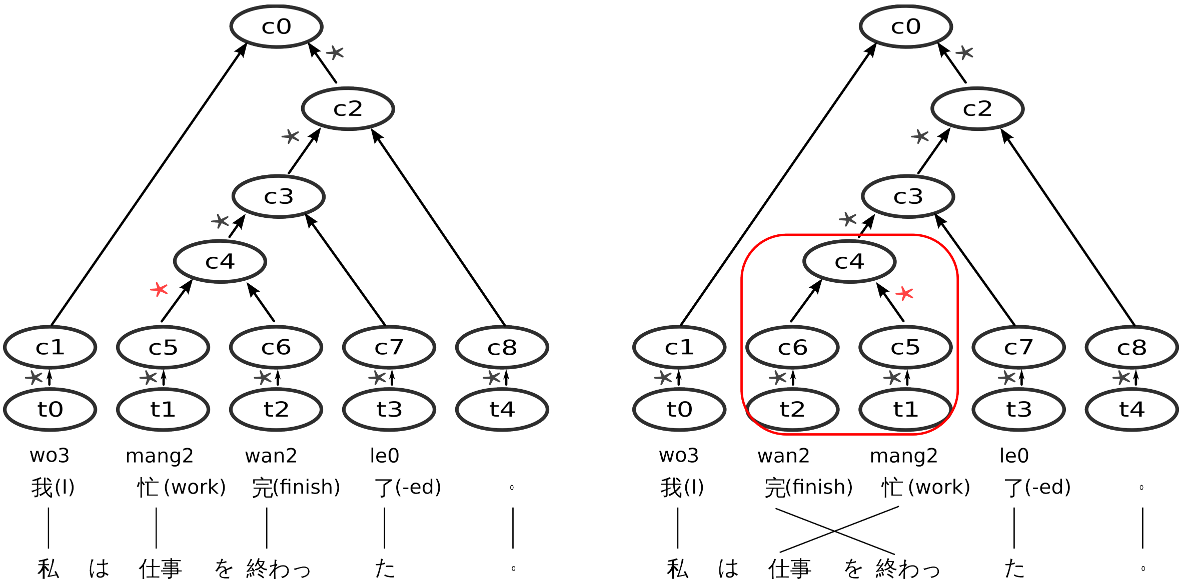
FIGURE 4.6: An example of showing the mis-reordering of a series of verbs in subordinate relationship while implementing HFC. Figure 4.6a give the original parse tree and its English translation. Figure 4.6b shows the wrongly reordered Chinese sentence along with its Japanese translations.

that acts as object of the first verb of “维持(maintain)”. Additionally, since HFC can only perform swapping operation on a binary tree, it might not be possible to obtain the optimal reordering due to constraints imposed by the binary structure, such as reordering between nodes of `c6` and `c8` in Figure 4.5b.

Figure 4.6 shows an example of two verbs, “可望(expect)” and “达到(reach)”, in a subordinate relationship. A verb phrase, which is composed by verb of “达到(reach)” and its object of “预期(anticipation)”, is the object of verb “可望(expect)”. However, since node of `c6` is not the syntactic head of node `c5` (See Figure 4.6a), HFC failed in reordering verb “可望(expect)” as shown in Figure 4.6b.

4.4.2 Complementizer

A complementizer is a particle used to introduce a complement. A very common complementizer in English is the word *that* when making a clausal complement. But in Chinese, it can denote other types of word, such as verbs, adjectives or quantifiers. Since the complementizer is identified as the dependent of the verb that it modifies in Chinese,



(A) Original HPSG Chinese parse tree (B) Reordered Chinese sentence

FIGURE 4.7: An example of showing the mis-reordering of complementizer while implementing HFC. Figure 4.7a shows the original parse tree and its English translation. Figure 4.7b shows the wrongly reordered Chinese sentence along with its Japanese translation.

whereas it is the syntactic head in Japanese, the inconsistency on head definition leads to undesired reordering result.

For instance, an original Chinese sentence in Figure 4.7a has already monotonically aligned with its Japanese translation. However, since “完(finish)” is a complementizer of “忙(work)” in Chinese, it has been swapped with its syntactic head, which produced a crossing alignment as shown in Figure 4.7b. It is challenging to normalize complementizer in Chinese and construct a single exception rule in HFC due to the multiformity of the source of complementizer.

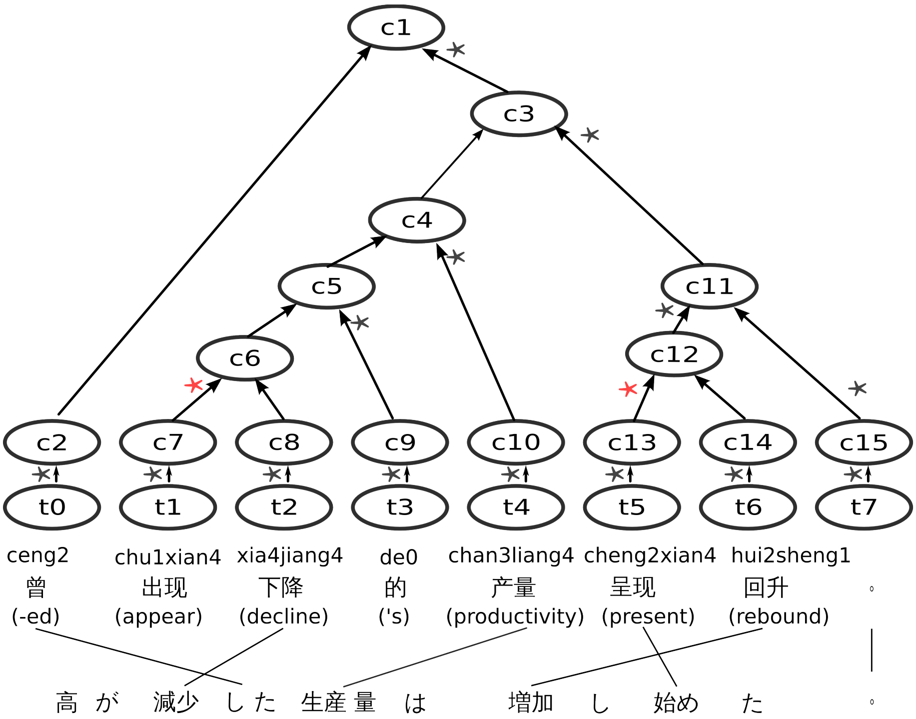
4.4.3 Adverbial modifier

There are many types of adverbial modifiers both in Chinese and Japanese. Unlike the adverb 不(not) that we discussed in Section 4.2.2, the ordinary adverbial modifier usually appears directly ahead of the verb that it modifies in both languages, but not in English. These adverbial modifiers are often recognized as dependents of their verbs, while the verbs are always recognized as syntactic heads. For this reason, adverbs and verbs will not preserve their relative position when they are reordered by using the Head Finalization principle.

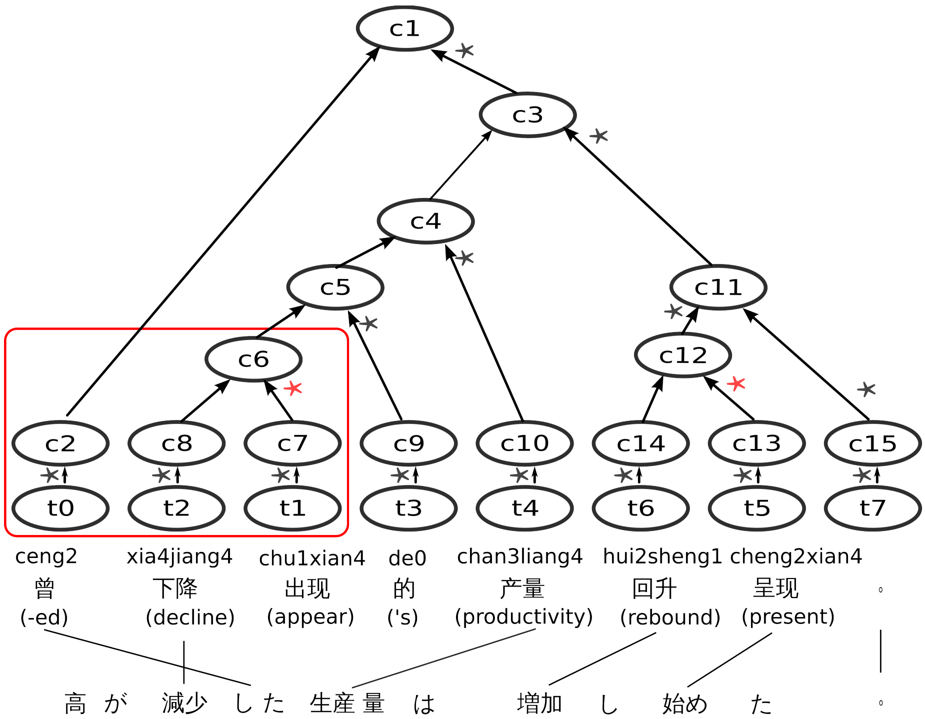
We illustrate this lack of coverage of HFC in Table 4.8. In the Chinese sentence, "曾(-ed)" is an adverb and it adds the past tense to the verb "下降(decline)" as an modifier. The word alignment with the Japanese translation shows that the adverbial modifier should be reordered to the right-hand side of its verb. However, since "曾(-ed)" is neither a syntactic head, nor in the same branch as its verb head "下降(decline)", it is impossible to reorder it to the desired position by using HFC reordering strategy of swapping branches. In other words, this problem is not derived solely by the fact that adverbs are not recognized as syntactic heads. In the hypothetical case that this adverb were considered as a syntactic head, HFC would incorrectly reordered it at the end of the sentence, while its optimal position would be to the right-hand side of its verb. As a result, the alignment of adverbial modifiers would still be non-monotonic. Meanwhile, "出现(appear)" and "下降(decline)" were swapped for the reason that both of them are verbs, which is one of HFC reordering issues and we have discussed in Section 4.4.1.

4.4.4 Verbal nominalization and nounal verbalization

Apart of identifying the relation among verbs in a verb serial, another difficulty of having little inflectional morphology in Chinese is to label POS tags for Chinese characters. The usages of some Chinese characters are extremely flexible, such as verb nominalization or noun verbalization, which appears frequently and commonly without any conjugation or declension. As a result, it is problematic to do disambiguation during POS tagging and parsing. For example, the Chinese character of "开发" has two regular syntactic functions without any inflection, as a verb means develop while as a noun means development. It is impractical to reliably tag without considering the context. In contrast, in Japanese, する is a symbol to identify verbs which can be also used as noun. For instance, "開発する(develop)" is the verb of the noun of "開発(development)". This ambiguity is prone to not only POS tagging error but also parsing error, and thus affects the identification of heads, which may lead to incorrect reordering.



(A) Original HPSG Chinese parse tree



(B) Reordered Chinese sentence

FIGURE 4.8: An example of showing the mis-reordering of adverbial modifier while implementing HFC. Figure 4.7a shows the original parse tree and its English translation. Figure 4.7b shows the wrongly reordered Chinese sentence along with its Japanese translation.

4.5 Summary

In this chapter, we have proposed a Chinese-to-Japanese reordering rules inspired by [1] based on a detailed linguistic analysis on Chinese HPSG and differences between Chinese and Japanese. The syntactic relationship of this language pair has not been carefully studied before in the machine translation field, and our work contributed in this direction.

The news experiment results show that, although a simple implementation of HF to reorder Chinese sentences performs better than baseline system, translation quality was substantially improved further by including linguistic knowledge into the refinement of the reordering rules. However, the experiment on patent domain did not show any advantage of either HF nor HFC.

In Section 4.4, we presented more patterns on reordering issues when reordering Chinese sentences to resemble Japanese word order. However, although HFC pre-reordering method has been used to reorder words better, it has shown its limitation on further improvement.

Chapter 5

Dependency Parsing based Pre-reordering for Chinese (DPC)

Chinese and Japanese have a different sentence structure. Reordering methods are effective, but need reliable parsers to extract the syntactic structure of the source sentences. However, Chinese has a loose word order, and Chinese parsers that extract the phrase structure do not perform well. In this chapter, we propose a block-based pre-reordering framework where only POS tags and unlabeled dependency parse trees are necessary, and linguistic knowledge on structural difference can be encoded in the form of reordering rules. We show significant improvements in translation quality of sentences from news domain, when compared to our HFC reordering methods.

5.1 Methodology

As we talked in Chapter 1, objects usually follow their verbs in SVO languages, but in Subject-Object-Verb (SOV) languages, objects precede them. Our objective is to reorder tokens in Chinese sentences (SVO) to resemble the word order of Japanese sentences (SOV). For that purpose, our method consists in moving verbs to the right-hand side of their objects. However, it is challenging to correctly identify the appropriate verbs and objects that trigger a reordering, and this section will be dedicated to that end.

More specifically, the first step of our method consists in identifying the appropriate verb (and certain tokens close to it) that need to be moved to the right-hand side of its object argument. The verb (and those accompanying tokens) will be moved as a block, preserving the relative order among them. We will refer to them as *Verbal block* (Vb). The second step will consist in identifying the right-most object argument of the verb under consideration, and moving the verbal block to the right-hand side of it. Finally, certain invariable grammatical particles in the original vicinity of the verb will also be reordered, but their positions will be decided relative to their verb.

In what follows, we describe in detail how to identify verbal blocks, their objects and the invariable grammatical particles that will play a role in our reordering method. The only information that will be used to perform the task will be the POS tags [1] of the tokens and their unlabeled dependency structures.

5.1.1 Identifying verbal block (Vb)

A verbal block (Vb) is composed of a head (Vb-H) and possibly accompanying dependents (Vb-D). In the Chinese sentence “我(I) 吃(eat) 了(-ed) 梨(pear).”, “吃” refers to the English verb “eat” and the aspect particle “了” adds a preterit tense to the verb. Tokens of “吃 了 (ate)” are an example of a Vb that should be reordered together without altering its inner word order, i.e. “我(I) 梨(pear) 吃(eat) 了(-ed).”, which matches SOV order in Japanese.

[1] We follow the POS tag guideline of the Penn Chinese Treebank v3.0. [92]

TABLE 5.1: Lists of POS tags in Chinese used to identify blocks of tokens to reorder (Vb-H, Vb-D, BEI lists), the POS tags of their dependents (RM-D lists) which indicate the reordering position, and other particles (Oth-DEP) that need to be reordered.

Category	POS tag Candidates
Vb-H	VV VE VC VA P
Vb-D	AD AS SP MSP CC VV VE VC VA
BEI	LB SB
RM-D	NN NR NT PN OD CD M FW CC
	ETC LC DEV DT JJ SP IJ ON
Oth-DEP	LB SB CS

Possible head of a verbal block (Vb-H) is verbs (tokens with POS tags of VV, VE, VC and VA), or preposition (token with POS tag of P). The Vb-H entry of Table 5.1 contains the list of POS tags for heads of verbal blocks. We use prepositions for Vb-H identification since they behave similarly to verbs in Chinese and should be moved to the right-most position in a prepositional phrase to resemble the Japanese word order. There are three conditions that a token should meet to be considered as a Vb-H:

i) Its POS tag is in the set of Vb-H in Table 5.1.

ii) It is a dependency head, which indicates that it may have an object as a dependent.

iii) It has no dependent whose POS tag is in the set of BEI in Table 5.1. BEI particles indicate that the verb is in passive voice and should not be reordered since it already resembles the Japanese order.

A bei-construction is a special structure that is commonly used to create passive voice in Chinese sentences. In order to compensate the lack of verb inflection in Chinese, particles are introduced to indicate the occurrence of a passive voice. These particles have a POS tag LB or SB, and are dependents of the verb. In Chinese, bei-constructions follow the OV word order, which is the same as the Japanese word order. That is, the verb is on the right-hand side of its object. For this reason, reordering is not required and we exclude Vb-H candidates that are involved in a bei-construction. Figure 5.1 illustrates this linguistic phenomenon. In the Chinese sentence, the main verb “批评(criticize)” is already on the right-hand side of its object, “学生(student)” in this case. This word

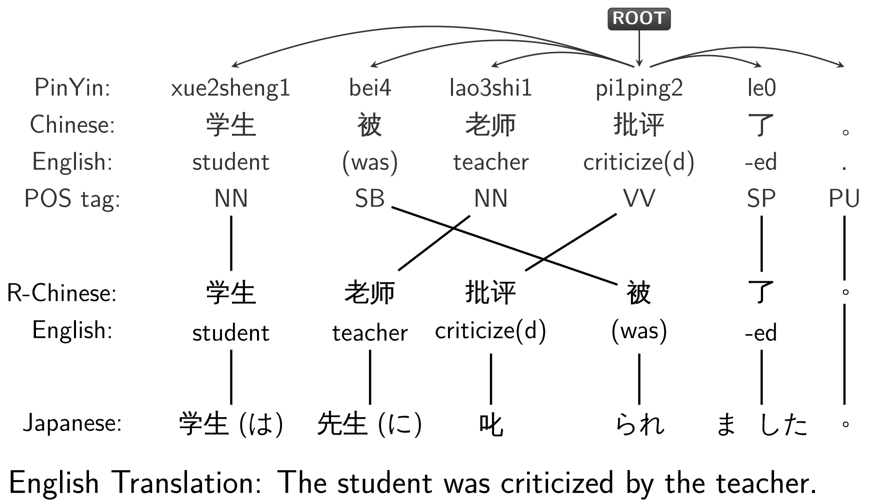


FIGURE 5.1: Example of bei-construction. R-Chinese shows the desired reordered Chinese.

order is the same as in the Japanese sentence. For this reason, no reorder is necessary between the main verb and its object, which motivates our condition on bei-constructions presented above.

Chinese language does not have inflection, conjugation, or case markers [103, 104]. For that reason, some adverbs (AD), aspect particles (AS) or sentence-final particles (SP) are used to signal modality, indicate grammatical tense or add tnoteA Chinese character is specially used to connect the verb phrase and its modifier. aspectual value to verbs. Words in this category preserve the order when translating to Japanese, and they will be candidates to be part of the verbal block (Vb-D) and accompany the verb when it is reordered. Other tokens in this category are coordinating conjunctions (CC) that connect multiple verbs, and both resultative 得(DER)[2] and manner 地(DEV)[3]. The full list of POS tags used to identify Vb-Ds can be found in Table 5.1. To be a Vb-D, there are three necessary conditions as well:

i) Its POS tag is in the Vb-D entry in Table 5.1.

ii) It is a dependent of a token that is already in the Vb.

iii) It is next to its dependency head or only a coordination conjunction is in between.

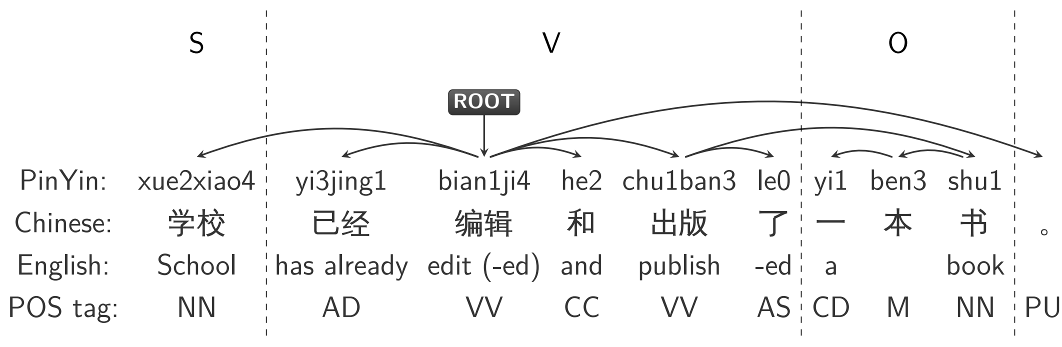
[2] A Chinese character is specially used between verbs or adjectives and their modifiers.
[3] A Chinese character is specially used to connect the verb phrase and its modifier.

To summarize, to build a verbal block (Vb), we first identify a token that meet the three Vb-H conditions. Then, we test the Vb-D conditions on tokens that are adjacent to the Vb-H and extend the verbal block to include qualified tokens as Vb-Ds. This process is iteratively applied to the adjacent tokens of a block until no more token can be added into the Vb, possibly nesting other verbal blocks if necessary.

Figure 5.2 shows an example of a dependency tree of a Chinese sentence that will be used to illustrate Vb identification. By observing the POS tags of the tokens in the sentence, only tokens of "编辑(edit)" and "出版(publish)" have the POS tag (i.e. VV) in the Vb-H entry of Table 5.1. Moreover, both tokens are dependency heads and do not have any dependent whose POS tag is in the BEI entry of Table 5.1. Thus, "编辑(edit)" and "出版(publish)" will be selected as Vb-H and form, by themselves, two separate incipient Vbs. We arbitrarily start building the Vb from the token of "出版(publish)", by analyzing its adjacent tokens that are its dependents.

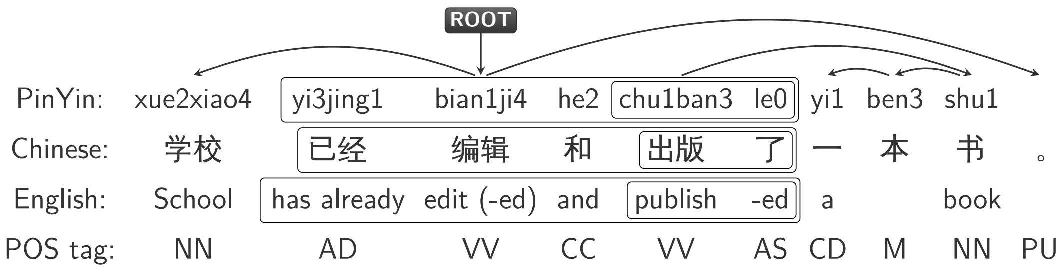
We observe that only "了(-ed)" is adjacent to "出版(publish)", it is its dependent, and its POS tag is in the Vb-D list of Table 5.1. Since "了(-ed)" meets all three conditions stated above, "了(-ed)" will be included in the Vb originated by "出版(publish)". The current Vb thus consists of the sequence of tokens "出版(publish)" and "了(-ed)", and the three conditions for Vb-D are tested on the adjacent tokens of this Vb. Since the adjacent tokens (or tokens separated by a coordinating conjunction) do not meet the conditions, the Vb is not further extended. Figure 5.2b shows the dependency tree where the Vb that consists of the tokens of "出版(publish)" and "了(-ed)" is represented by a rectangular box.

By checking in the same way, there are three dependents that meet the requirements of being Vb-Ds for "编辑(edit)": "已经(has already)", "和(and)" and "出版(publish)" and hence this Vb consists of three tokens and one Vb. The outer rectangular box in Figure 5.2b shows that the Vb with "编辑(edit)" as the Vb-H. Nested Vbs are merged and reordered as one in the end. Figure 5.2c shows an image of how the merged Vb will be reordered while the inner orders are kept. Note that the order of building Vbs from which Vb-H, "出版(publish)" or "编辑(edit)", will not affect any change of the final result.

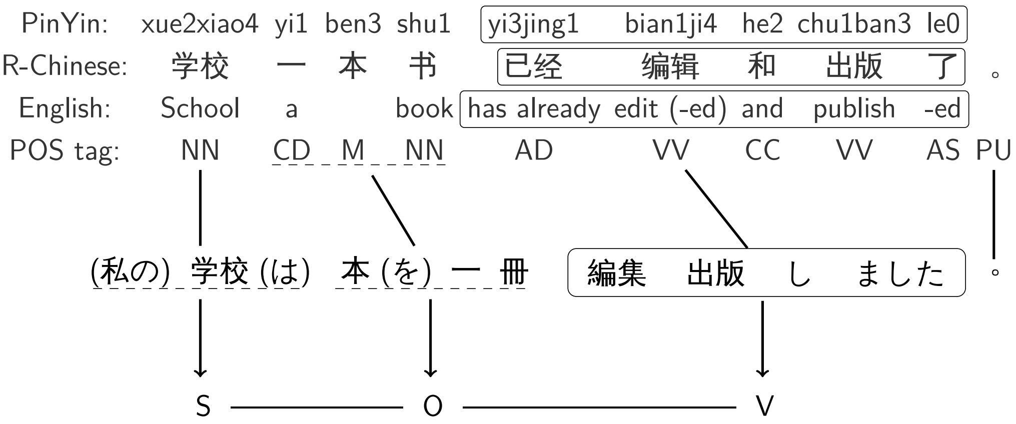


English Translation: (My) school has already edited and published a book.

(A) Original dependency tree



(B) Vbs in rectangular boxes



(C) Merged and reordered Vb

FIGURE 5.2: An example for showing how to detect and reorder a Vb and reordering a Chinese SVO sentence to be a Japanese SOV word order. In each subfigure, Chinese Pinyin, Chinese token, and token-to-token English translation are listed in three lines. POS tag of each Chinese token are also given in the first two subfigures. In Figure 5.2c, word alignment between reordered Chinese sentence and its Japanese counterpart is given as well.

5.1.2 Identifying the right-most object dependent (RM-D)

In the most general form, objects are dependents of a verbal block[4] that act as its arguments. While the simplest objects are nouns (N) or pronouns (PN), they can also be comprised of noun phrases or clauses [105], such as nominal groups, finite clauses (e.g. *that* clauses, *wh*-clauses) or non-finite clauses (e.g. *-ing* clauses), among others.

For every Vb in a verb phrase, clause, or sentence, we define the right-most object dependent (RM-D) as the word that:

i) its POS tag is in the RM-D entry of Table 5.1,

ii) its dependency head is inside of the verbal block, and

iii) is the right-most one among all dependents of the verbal block that satisfy the first two conditions.

All verbal blocks in the phrase, clause, or sentence will move to the right-hand side of their correspondent RM-Ds recursively. Note that, RM-D is not necessarily to be the right-most dependent of the Vb in a dependency tree. As an illustration that the punctuation of period will not be recognized as the RM-D of its dependency head by reason of not a part of the object, even though it is in the right most position of a dependency tree.

Figure 5.2b and Figure 5.2c show a basic example of object identification. The Chinese word corresponding to "书(book)" is a dependent of a token within the verbal block and its POS tag belongs to the RM-D entry list of Table 5.1, namely NN. Since it is the only qualified dependent of the verbal block, "书(book)" is identified as the right-most object dependent of the Vb, and the Vb will move to the right-hand side of it to resemble the Japanese word order.

A slightly more complex example can be found in Figure 5.3. In this example, there is a coordination structure of verb phrases, and the dependency tree shows that the first verb, "吃(eat)", appears as the dependency head of the second verb, "去(go)". The direct right-most object dependent (RM-D) of the first verb, "吃(eat)", is the word "午

[4]Dependents of a verbal block are dependents of any token within the verbal block.

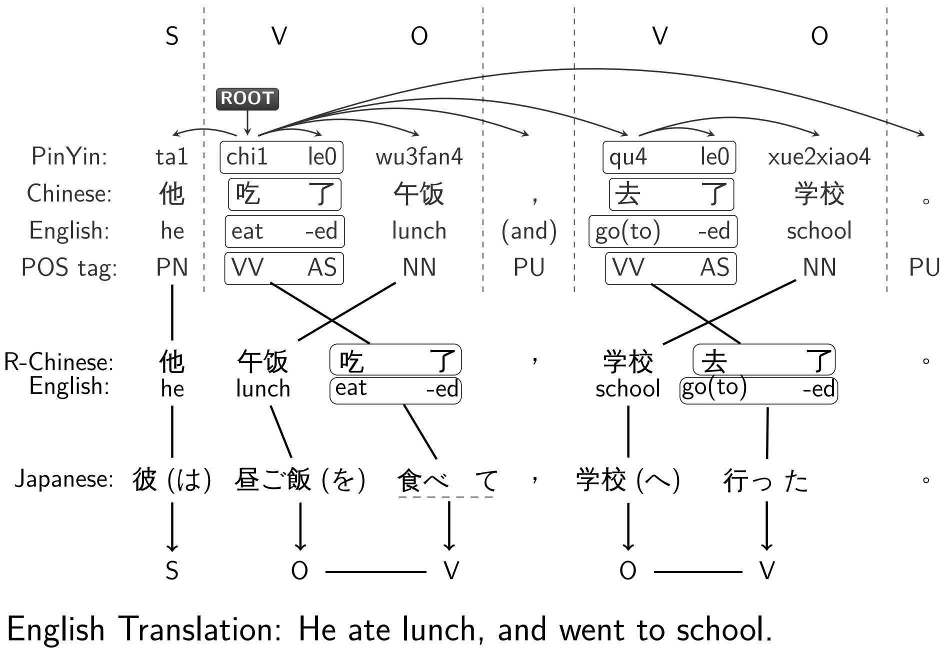


Figure 5.3: An example of a Chinese sentence with a coordination of verb phrases as predicate. Subject(S), verbs(V), and objects(O) are displayed for both verb phrases and the Japanese translation. Alignment lines between the original Chinese sentence and the reordered Chinese sentence indicate the reordering trace of verbal blocks(Vb) while alignment between reordered Chinese and its Japanese counterpart shows the reordering result.

饭(lunch)”, and the verb “吃(eat)” will be moved to the right-hand side of its RM-D. Meanwhile, there are two verbal blocks that have been composed, i.e., “吃(eat) 了(-ed)” and “去(go to) 了(-ed)”. The R-Chinese shows the reordered Chinese.

There are cases, however, where there is no coordination structure of verb phrases but a similar dependency relation occurs between two verbs. Figure 5.4 illustrates one of these cases, where the main verb “鼓励(encourage)” has no direct dependent that can be considered as an object since no direct dependent has a POS tag in the RM-D entry of Table 5.1. Instead, an embedded clause (SVO) appears as the object argument of the main verb, and the main verb “鼓励(encourage)” appears as the dependency head of the verb “参与(participate)”.

In the news domain, reported speech is a frequent example that follows this pattern. In our method, if the main verb of the sentence (labeled as ROOT) has dependents but none of them is a direct object, we move the main verb to the end of the sentence. As for the embedded clause “学生(student) 参与(participate) 社会(social) 实践(practice)”,

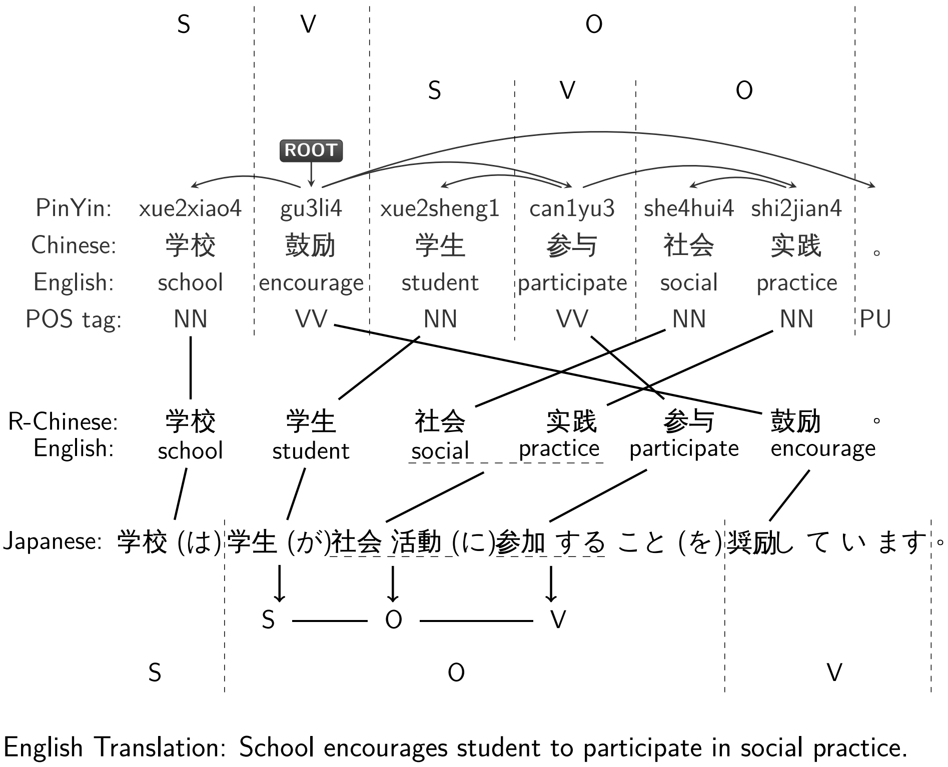


FIGURE 5.4: An example of a Chinese sentence of reported speech that a subordinate clause as object. Subject(S), verbs(V), and objects(O) are displayed for the main clause, the subordinate clause and the Japanese sentence. Alignment lines between the original Chinese sentence and the reordered Chinese sentence indicate the reordering trace of verbal blocks(Vb). Reordered Chinese is aligned with its Japanese translation.

the verbal block of the clause is the word "参与(participate)" and its object is "实践(practice)". Applying our reordering method, the clause order results in "学生(student) 社会(social) 实践(practice) 参与(participate)". The result is an SOV sentence with an SOV clause, which resembles the Japanese word order.

5.1.3 Identifying other particles (Oth-DEP)

In Chinese, certain invariable grammatical particles that accompany verbal heads have a different word order relative to their heads, when compared to Japanese. Those particles are typically 被(passive) particles, which have their POS tags as LB or SB, and subordinating conjunctions which have POS tag as CS. Those particles appear on the left-hand side of their dependency heads in Chinese, and they should be moved to the right-hand side of their dependency heads for them to resemble the Japanese word order. Reordering these particles in our framework can be summarized as:

i) Find dependents of a verbal head (Vb-H) whose POS tags are in the Oth-DEP entry of Table 5.1.

ii) Move those particles to the right-hand side of their (possibly reordered) heads.

iii) If there is more than one such particle, move them keeping the relative order among them.

5.1.4 Summary of the reordering framework

Based on the definitions above, our dependency parsing based pre-reordering framework can be summarized in the following steps:

1. Obtain POS tags and an unlabeled dependency tree of a Chinese sentence.
2. Obtain reordering candidates: Vbs.
3. Obtain the right-most object dependent (RM-D) of each Vb.
4. Reorder each Vb in two exclusive cases by following the order:
   (a) If RM-D exists, reorder Vb to be the right-hand side of RM-D.
   (b) If Vb-H is ROOT and its RM-D does not exist, reorder Vb to the end of the sentence.
   (c) If none of above two conditions is met, no reordering happens.
5. Reorder grammatical particles (Oth-DEPs) to the right-hand side of their corresponding Vbs.

Note that, unlike other works in reordering distant languages [1, 2, 63], we do not prevent chunks from crossing punctuations or coordination structures. Thus, our method allows to achieve an authentic global reordering in reported speech, which is an important reordering issue in news domains.

In order to illustrate our method, a more complicated Chinese sentence example is given in Figure 5.5, which includes the unlabeled dependency parse tree of the original Chinese sentence, and the word alignment between reordered Chinese sentence and its Japanese counterpart, etc.

Based on both POS tags and the unlabeled dependency tree, first step of our method is to obtain all Vbs. For all heads in the tree, according to the definition of Vb introduced in Section 5.1.1, there are six tokens which will be recognized as the candidates of Vb-Hs, that is “报道(report)”, “随着(with)”, “进入(enter)”, “成为(become)”, “加强(strengthen)”, and “力促(urge)”. Then, for each of the candidate, its direct dependents will be checked if they are Vb-Ds. For instance, for the verb of “进入(enter)”, its dependents of “逐渐(gradually)” and “了(-ed)” will be considered as the Vb-Ds. For the case of “加强(strengthen)”, instead of being a Vb-H, it will be recognized as Vb-D of the Vb “力促(urge)” since it is one of the direct dependents of “力促(urge)” with a qualified POS tag for Vb-D. Therefore, there are five Vbs in total, which are “报道(report)”, “随着(with)”, “逐渐(gradually) 进入(enter) 了(-ed)”, “成为(become)”, and “加强(strengthen) 力促(urge)”.

The next step is to identify RM-D for each Vb, if there is one. By checking all conditions, four Vbs have their RM-Ds: “发展(development)” is the RM-D of the Vb “随着(with)”; “中国(China)” is the RM-D of the Vb “逐渐(gradually) 进入(enter) 了(-ed)”; “节日(festival)” is the RM-D of the Vb “成为(become)”; “买气(purchase)” is the RM-D of the Vb “加强(strengthen) 力促(urge)”.

After obtaining all RM-Ds, we find those Vbs that have RM-Ds and move them to right of their RM-Ds. As for the case of “报道(report)”, since it is the root and does not have any matched RM-D, it will be moved to the end of the sentence, before any final punctuation. Finally, since there is no any invariable grammatical particle in the sentence that need to be reordered, reordering has been finished. From the alignments between the reordered Chinese and its Japanese translation showed in the figure, an almost monotonic word alignment has been achieved.

For comparison purposes, particle seed words had been inserted into the reordered sentences in the same way as the HFC method, which is using the information of predicate argument structure output by *Chinese Enju* [40]. We therefore can not entirely disclaim

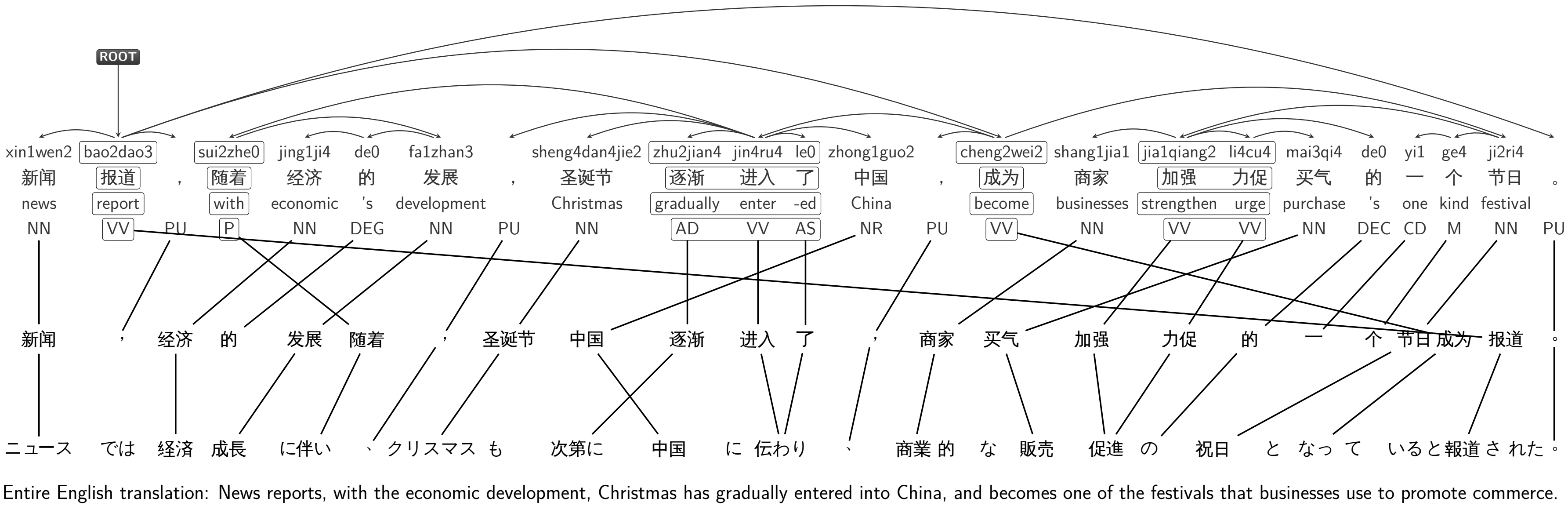


Figure 5.5: Dependency parse tree of a complex Chinese sentence example, and word alignments for reordered sentence with its Japanese counterpart. The first four lines are Chinese Pinyin, tokens, word-to-word English translations, and the POS tags of each Chinese token. The fifth line shows the reordered Chinese sentence while the sixth line is the segmented Japanese translation. The entire English translation for the sentence is showed in the last line.

the use of the HPSG parser at the present stage in our method. However, we believe that dependency parser can provide enough information for inserting particles.

5.2 Evaluation

5.2.1 Experiment setting

In order to compare the performance with our HFC, we evaluate DPC with the similar prepossessed corpora from both news domain and patent domain, which were described in Section 4.3.1. However, after extracting the POS tags of Chinese from the result of *Berkeley parser*, we use *Corbit* [41] to obtain the dependency structure of Chinese sentences for operating DPC. In addition, baseline system and evaluation metrics are exactly the same as described in Section 4.3.1.

5.2.2 Results

In phrase-based machine translation, the distortion limit is a parameter that governs the maximum distance at which a phrase can be translated. In statistical machine translation between language pairs with a different sentence structure, large distortion limits might be appropriate to allow translating phrases that align at the beginning and at the end of Chinese and Japanese sentences, respectively.

In Tables 5.3, 5.4 and Table 5.5, 5.6, we present the results with different size of training data for both news and patent corpora respectively, and we mark in bold the result for the optimal distortion limit. In general, the patent domain requires larger distortion limits than news domain, which might be due to sentences in patent domain being longer (in average) than sentences in news domain.

DPC pre-reordering method shows a consistent superior performance in terms of BLEU and RIBES (precision metrics), and WER, PER and TER (error metrics), for all distortion limits in both domains. DPC achieves the best performance using distortion limits that range between 6 and 12 in the news domain, and in the range between 9 and 12 in the patent domain.

Considering only the results of every method for their optimal distortion limit, we find small or inconsistent differences in performance between the baseline, HF and HFC methods in terms of our performance metrics. In the news domain, DPC obtains small improvements in BLEU and RIBES over the second best performing system, but slightly larger improvements in terms of WER, PER and TER. In the patent domain, however, DPC obtains large improvements with respect to HFC in terms of average on *BLEU* (2.9 and 3.6 points), average on *RIBES* (2.2 and 2.3 points), average on *WER* (5.3 and 5.5 points), average on *PER* (4.4 and 4.3 points), and TER (3.8 and 3.8 points).

Table 5.2 shows two examples from the development data set that compares qualitatively the reordering capabilities of HFC and DPC. In the first example (top), HFC produces an inverted constituent order (3, 2 and 1) with respect to the Japanese reference (1, 2 and 3), which is undesirable. In this example, DPC correctly follows the Japanese constituent order, except for the Chinese word $_1$使用(use), which is equivalent to the Japanese 用いて and that is incorrectly splitted from its constituent and placed wrongly between constituents 2 and 3. The reason is that the parse incorrectly recognized $_2$来(to) 分析(analyze) as the modifies of $_1$使用(use). In the second example (bottom), HFC produces an incorrect constituent order (3, 1 and 2) with respect to the Japanese reference (1, 2 and 3). The reason is that as a modality verb $_3$能够(can) is not recognized as the head. Meanwhile, DPC produces an even worse constituent order, where the Chinese equivalent to the Japanese constituent 1 is splitted into three parts, and constituents 2 and 3 are placed inside of the parenthesis. Such incorrect word order produced by DPC was caused by the dependency parser mis-recognizing $_1$健全(soundness) as the main verb of the sentence (and hence, DPC moved it to the end of the sentence), while $_2$作为(be used as) $_3$能够(can) where wrongly recognized as modifiers of $_1$健全(soundness).

TABLE 5.2: Reordering Examples of HFC and DPC from the development set. Constituents are underlined and indexed. Verbs and particles of interest appear in red, and their constituents appear in purple.

Japanese-1	この よう な 配列 類似 性 は 、 例えば 、 1前述 の よう な FASTA 等 の プログラム を 用い て 2算出 する こと が 3できる 。
Original Chinese-1 (English)	此种(This) 类型(type) 的 序列(sequence) 相似性(similarity) 3可(can) 1使用(use) 诸如(such as) 上面(above) FASTA 的 程序(program) 2来(to) 分析(analyze) 。
HFC-1	此种 类型 的 序列 相似性 3可 2来 分析 1诸如 上面 FASTA 的 程序 使用 。
DPC-1	此种 类型 的 序列 相似性 1诸如 上面 FASTA 的 程序 2来 分析 1使用 3可 。
Japanese-2	物体 中 の 音速 は 、 当該 物体 の 弾性 的 な 性質 によって 変化 する もの で ある から 、 骨 中 の 音速 を 測定 する こと で 骨 強度 1（ 骨 の 健全 性 ） の 指標 と 2する こと が 3できる 。
Original Chinese-2 (English)	物体(object) 中(in) 的(of) 声速(velocity of sound) 根据(depend on) 该(the) 物体(object) 的(of) 弹性(elasticity) 的(of) 性质(property) 而 变化(change) 发生(happen) , 所以(so) 通过(by) 测定(measure) 骨中(in bone) 的(of) 声速 (velocity of sound) 3能够(can) 2作为(be used as) 骨(bone) 强度(strength) 1(骨(bone) 的(of) 健全(soundness) 性(-ness)) 的(of) 指标(indicator) 。
HFC-2	物体 中的 声速 该 物体 的 弹性 的 性质 根据 而 变化 发生 , 所以 测定 骨 中 的 声速 通过 3能够 骨 强度 1(骨 的 健全 性) 的 指标 2作为 。
DPC-2	物体 中的 声速 该 物体 的 弹性 的 性质 根据 变化 而 发生 , 所以 测定 骨 中 的 声速 通过 骨 强度 1骨 (2作为 3能够 1的 性) 的 指标 1健全 。

TABLE 5.3: Evaluation of translation quality on news domain (Training 1). Results are given in terms of BLEU, RIBES, WER, PER, and TER for baseline, HF, HFC and DPC along with different values of distortion limit (dl). Results with a confidence over 95% are marked with superscripts. Superscript 1 denotes the system is significantly better than baseline. Superscript 2 denotes the system is significantly better than HF. Superscript 3 denotes the system is significantly better than HFC.

	dl	0	1	2	3	4	5	6	7	9	12
BLEU	Baseline	38.29	38.58	38.55	38.72	38.91	39.11	39.16	39.21	**39.47**	39.44
	HF	39.09^{1}	39.05^{1}	39.22^{1}	39.20^{1}	39.34^{1}	$\mathbf{39.59}^{1}$	39.53	39.57	39.30	39.54
	HFC	39.55^{12}	39.48^{12}	39.00^{1}	39.62^{12}	39.70^{12}	39.49	39.66^{1}	$\mathbf{39.79}^{1}$	39.66^{2}	39.66
	DPC	39.59^{12}	39.66^{12}	39.62^{123}	39.77^{12}	39.68^{1}	39.43	39.94^{12}	39.87^{1}	$\mathbf{40.14}^{123}$	39.85^{1}
RIBES	Baseline	84.55	84.59	84.59	84.60	84.87	84.80	**85.07**	84.95	84.91	84.87
	HF	84.77	84.79	84.77	84.85	84.66	84.92	84.90	84.90	**84.95**	84.89
	HFC	84.86^{1}	84.94^{1}	84.91^{1}	85.01^{1}	84.98^{2}	84.77	85.08	84.99	**85.09**	85.04
	DPC	85.07^{123}	85.12^{12}	85.13^{12}	85.22^{12}	85.11^{2}	85.18^{123}	$\mathbf{85.30}^{2}$	85.25^{123}	85.29^{12}	85.29^{123}
WER	Baseline	51.93	51.68	51.83	51.38	**51.08**	50.84	50.68	50.53	50.61	51.15
	HF	50.48	50.67	50.50	50.31	50.44	**50.01**	50.27	50.16	50.38	50.74
	HFC	50.00	50.03	50.20	49.91	49.79	50.23	49.72	**49.68**	49.70	49.92
	DPC	49.28	49.26	49.24	49.07	49.22	49.39	**48.74**	48.85	48.86	48.83
PER	Baseline	31.52	31.27	31.23	31.38	31.11	31.04	31.14	31.10	30.88	**30.81**
	HF	30.82	30.81	30.71	30.68	30.54	30.58	30.61	30.68	30.73	**30.34**
	HFC	30.65	30.65	31.00	30.52	**30.46**	30.92	30.51	30.66	30.54	30.59
	DPC	30.52	30.50	30.47	30.48	30.46	30.57	30.39	30.36	**30.25**	30.58
TER	Baseline	48.11	47.79	47.80	47.56	47.09	46.89	46.90	46.81	**46.65**	46.89
	HF	46.44	46.59	46.43	46.21	46.37	45.93	46.16	**46.12**	46.14	46.16
	HFC	45.99	46.02	46.16	45.89	45.67	46.21	45.71	45.75	**45.62**	45.73
	DPC	45.87	45.82	45.81	45.65	45.81	45.79	45.41	45.49	**45.34**	45.58

TABLE 5.4: Evaluation of translation quality on news domain (Training 2). Results are given in terms of BLEU, RIBES, WER, PER, and TER for baseline, HF, HFC and DPC along with different values of distortion limit (dl). Results with a confidence over 95% are marked with superscripts. Superscript 1 denotes the system is significantly better than baseline. Superscript 2 denotes the system is significantly better than HF. Superscript 3 denotes the system is significantly better than HFC.

	dl	0	1	2	3	4	5	6	7	9	12
BLEU	Baseline	38.35	38.20	38.32	38.63	38.81	39.21	39.20	**39.43**	39.41	39.20
	HF	39.15^{1}	39.48^{1}	36.86	39.66^{1}	39.41^{1}	39.70^{1}	39.55	36.29	**40.00**^{1}	39.85^{1}
	HFC	39.54^{12}	39.44^{1}	39.61^{12}	39.48^{1}	37.39	39.65^{1}	39.69^{1}	39.79^{2}	39.91^{1}	**39.94**^{1}
	DPC	39.62^{12}	39.44^{1}	39.56^{12}	39.70^{1}	39.66^{13}	39.75^{1}	39.82^{1}	**40.01**^{12}	39.95^{1}	39.81^{1}
RIBES	Baseline	84.53	84.60	84.64	84.66	84.65	85.00	85.10	85.10	**85.12**	84.83
	HF	84.84^{1}	84.78	84.06	84.85	84.80	84.96	84.80	82.62	**85.02**	84.71
	HFC	84.92^{1}	84.77	84.99^{12}	84.79	84.42	84.92	84.91	84.88^{2}	**85.10**	84.94^{2}
	DPC	85.17^{123}	84.94^{1}	85.19^{12}	85.14^{123}	85.23^{123}	85.25^{123}	85.26^{23}	85.18^{23}	85.21	**85.27**^{123}
WER	Baseline	52.07	52.15	52.02	51.59	51.39	50.85	50.75	50.42	**50.35**	51.00
	HF	50.59	50.46	52.62	50.18	50.19	50.04	50.10	53.93	**49.84**	51.02
	HFC	50.10	50.25	49.99	50.18	51.55	50.03	50.03	49.99	**49.64**	49.96
	DPC	49.18	49.66	49.17	49.32	49.00	49.15	**48.94**	48.99	48.99	48.95
PER	Baseline	31.39	31.48	31.40	31.00	31.31	30.85	30.87	30.83	**30.80**	30.91
	HF	30.59	30.48	31.31	30.45	30.47	30.45	30.45	30.73	**30.24**	30.37
	HFC	30.52	30.62	30.36	30.62	31.43	30.48	30.51	**30.34**	**30.34**	30.40
	DPC	30.50	30.55	30.51	30.35	30.30	30.46	30.31	30.21	**30.07**	30.28
TER	Baseline	48.09	48.12	47.97	47.40	47.60	46.94	46.84	46.74	**46.53**	46.73
	HF	46.42	46.35	48.22	46.09	46.00	45.92	45.94	49.10	**45.69**	46.24
	HFC	46.00	46.20	45.88	46.14	47.27	45.85	45.94	45.80	**45.63**	45.78
	DPC	45.83	46.29	45.80	45.92	45.56	45.88	45.61	45.70	45.54	**45.47**

TABLE 5.5: Evaluation of translation quality on patent domain (Training 1). Results are given in terms of BLEU, RIBES, WER, PER, and TER for baseline, HF, HFC and DPC along with different values of distortion limit (dl). Results with a confidence over 95% are marked with superscripts. Superscript 1 denotes the system is significantly better than baseline. Superscript 2 denotes the system is significantly better than HF. Superscript 3 denotes the system is significantly better than HFC.

	dl	0	1	2	3	4	5	6	7	9	12
BLEU	Baseline	45.51	45.69	45.64	45.95	46.50	46.79	47.02	47.68	48.01	**48.03**
	HF	44.97	45.44	45.51	46.61	46.80	47.59^{1}	48.00^{1}	48.52^{1}	**49.11**^{1}	48.25
	HFC	45.06	45.04	45.27	45.45	45.45	47.20	48.20^{1}	48.49^{1}	48.22	**48.84**^{12}
	DPC	48.06^{123}	48.06^{123}	48.23^{123}	48.88^{123}	49.23^{123}	50.11^{123}	50.20^{123}	50.89^{123}	51.31^{123}	**51.44**^{123}
RIBES	Baseline	84.32	84.55	84.37	84.38	84.82	84.73	84.88	85.11	**85.41**	84.93
	HF	84.19	84.33	84.29	84.81	84.81	85.06	85.15	85.30	**85.51**	85.09
	HFC	84.16	84.17	84.21	84.19	84.19	84.82	85.18	**85.38**	85.18	85.20
	DPC	86.49^{123}	86.46^{123}	86.59^{123}	86.65^{123}	86.89^{123}	86.92^{123}	87.14^{123}	87.14^{123}	87.29^{123}	**87.32**^{123}
WER	Baseline	49.03	48.80	49.01	48.58	47.92	47.87	47.90	47.36	**47.32**	48.85
	HF	50.70	50.04	50.41	48.93	49.17	48.36	47.88	47.42	**47.24**	49.42
	HFC	50.40	50.52	50.36	50.17	50.17	48.58	47.33	**47.09**	47.85	48.37
	DPC	44.79	44.98	44.66	44.20	43.83	43.31	43.22	42.84	**42.83**	43.30
PER	Baseline	26.41	26.37	26.30	26.02	25.74	25.37	25.06	24.93	24.61	**24.36**
	HF	27.97	27.39	27.89	26.71	26.99	26.67	26.26	**25.61**	**25.61**	26.63
	HFC	28.16	28.11	28.06	28.24	28.24	26.57	25.68	**25.29**	25.87	25.83
	DPC	22.88	23.36	22.95	22.97	22.99	22.52	22.35	22.16	21.93	**21.57**
TER	Baseline	41.67	41.72	41.68	41.09	40.57	40.10	40.11	39.65	**39.40**	40.08
	HF	42.34	41.82	42.19	40.93	40.93	40.18	39.68	39.07	**38.84**	40.18
	HFC	42.36	42.37	42.21	42.08	42.08	40.35	39.34	**38.97**	39.51	39.57
	DPC	38.15	38.49	37.95	37.83	37.44	36.66	36.53	36.12	36.04	**35.79**

Table 5.6: Evaluation of translation quality on patent domain (Training 2). Results are given in terms of BLEU, RIBES, WER, PER, and TER for baseline, HF, HFC and DPC along with different values of distortion limit (dl). Results with a confidence over 95% are marked with superscripts. Superscript 1 denotes the system is significantly better than baseline. Superscript 2 denotes the system is significantly better than HF. Superscript 3 denotes the system is significantly better than HFC.

	dl	0	1	2	3	4	5	6	7	9	12
BLEU	Baseline	50.83	50.59	51.30	51.60	51.74	52.34	52.92	53.40	54.00	**54.51**
	HF	51.53	51.81^{1}	51.74	52.38^{1}	53.70^{1}	54.05^{1}	54.17^{1}	54.82^{1}	**55.28**^{1}	55.22
	HFC	51.77^{1}	51.27	51.73	51.93	52.93^{1}	53.75^{1}	53.83^{1}	54.05	55.03^{1}	**56.13**^{12}
	DPC	54.76^{123}	54.80^{123}	54.93^{123}	55.72^{123}	56.48^{123}	57.31^{123}	57.91^{123}	58.30^{123}	59.01^{123}	**59.18**^{123}
RIBES	Baseline	85.87	85.78	86.03	86.06	86.08	86.31	86.66	86.96	**87.18**	87.16
	HF	85.79	85.95	85.86	86.18	86.48	86.57	86.68	86.76	86.88	**86.92**
	HFC	86.03	85.85	86.04	85.91	86.30	86.55	86.62	86.55	86.98	**87.34**^{2}
	DPC	88.16^{123}	88.21^{123}	88.22^{123}	88.34^{123}	88.75^{123}	88.88^{123}	88.97^{123}	88.93^{123}	**89.25**^{123}	89.23^{123}
WER	Baseline	45.37	45.26	44.68	44.51	44.39	43.49	43.03	42.49	42.40	**42.29**
	HF	45.59	45.09	45.45	44.76	43.67	43.40	43.46	42.95	**42.73**	43.58
	HFC	45.06	45.59	45.09	45.05	44.17	43.49	43.63	43.25	42.50	**41.63**
	DPC	40.17	40.14	39.90	39.32	38.43	38.01	37.46	37.46	**36.66**	37.24
PER	Baseline	23.71	24.00	23.54	23.24	23.17	22.91	22.77	22.43	21.96	**21.74**
	HF	24.78	24.38	24.82	24.71	23.51	23.66	23.94	23.44	**23.23**	23.31
	HFC	24.36	24.92	24.26	24.39	23.76	23.53	23.57	23.99	23.07	**21.52**
	DPC	20.11	20.01	20.06	19.98	19.94	19.18	18.98	18.90	18.55	**18.51**
TER	Baseline	37.74	37.67	37.32	37.03	36.86	36.20	35.93	35.45	35.04	**34.60**
	HF	37.58	37.23	37.50	37.13	35.72	35.53	35.55	35.06	**34.73**	35.08
	HFC	37.15	37.62	37.17	37.09	36.11	35.44	35.53	35.52	34.55	**33.31**
	DPC	33.57	33.49	33.47	33.10	32.32	31.78	31.18	31.31	**30.56**	30.65

5.3 Summary

In the chapter, we have analyzed the differences in word order between Chinese and Japanese sentences. We captured the regularities of ordering differences between Chinese and Japanese sentences, and proposed a framework to reorder Chinese sentences to resemble the word order of Japanese.

Our framework consists in three steps. First, we identify verbal blocks, which consist of Chinese words that will move all together as a block without altering their relative inner order. Second, we identify the right-most object of the verbal block, and move the verbal block to the right of it. Finally, we identify invariable grammatical particles in the original vicinity of the verbal block and move them relative to their dependency heads.

Our framework only uses the unlabeled dependency structure of sentences and POS tag information of words. We compared our system to a baseline phrase-based SMT system and our head finalization Chinese system. Our DPC method obtained a Chinese word order that is more similar to Japanese word order, and we showed its positive impact on translation quality.

Chapter 6

Effects of Parsing Errors on Pre-reordering

Linguistically motivated reordering methods have been developed to improve word alignment especially for statistical machine translation on long distance language pairs as we described in previous chapters. However, since they highly rely on the parsing accuracy, it is useful to explore the relationship between parsing and reordering. In this chapter, for Chinese-to-Japanese SMT, we carry out a three-stage incremental comparative analysis to observe the effects of different parsing errors on reordering performance by combining empirical and descriptive approaches. For the empirical approach, we quantify the distribution of general parsing errors along with reordering qualities whereas for the descriptive approach, we extract seven influential error patterns and examine their correlation with reordering errors.

6.1 Analysis Method

Although in Section 4.4 of Chapter 4, we introduced several unsolved pre-reordering issues for HFC including parsing errors, we did not quantify the effects of parsing errors on the pre-reordering method. In order to draw a better image of how much the parsing errors influence pre-reordering methods do, we combine an empirical approach with a descriptive approach to observe the effects of parsing errors on pre-reordering performance in three stages: preliminary experiment stage, POS tag level stage, and dependency type level stage. First, we provide a general idea of the sensitiveness of parsing errors on both HFC and DPC reordering methods. Then, we focus on DPC and use POS tags to identify parsing errors and quantify the aggregate impact on DPC reordering performance since it is important and useful to help us to further improve our DPC pre-reordering method. Finally, we go a step further on the analysis for DPC that we define several concrete error patterns and examine their effects on DPC reordering qualities.

In order to test for an upper bound of the reordering performance and examine the specific parsing errors that affect reordering, one way is to contrast the reordering based on error-free parse trees, which are considered as Gold-Trees, with the reordering based on parse trees that are generated by parsers, which are referred as Auto-Trees in the following analysis.

In the preliminary experiment stage, we set up two benchmarks in two scenarios for both HFC and DPC. In scenario 1, the benchmark is manually reordered Chinese sentence on the basis of Japanese reference. By measuring the word order similarities between the benchmark and the Gold-Tree based reordered sentence as well as between the benchmark and the Auto-Tree based reordered sentence separately, we quantify the extent of parsing errors that influence reordering. Meanwhile, the former measurement shows additionally the general figure of the upper bound of the reordering method. However, since it is not only time-consuming but also labor-intensive to set up the benchmark in scenario 1, we use the Japanese reference as the benchmark in scenario 2 and follow the same strategies as in scenario 1 to calculate the word order similarities. More detailed description on the preliminary experiment is given in Section 6.2.

Starting from POS tag level stage, we focus of DPC only. We compare the Gold-tree with Auto-Tree along with reordering quality to explore the relationship between general parsing errors and reordering from two aspects: the percentages of top three most frequent dependent's POS tags that point to wrong heads and the percentages of top two most frequent head's POS tags that are recognized wrongly. The percentages of other POS tags are not provided because they are negligible. Our objective is to profile general parsing errors' distribution. However, this does not imply that those errors are the cause of the reordering errors. Section 6.3.1 includes more concrete analysis results.

In dependency type level stage, we classify the most influential dependency parsing errors on reordering into three super-classes and seven sub-classes according to the methodology of the reordering method. We then plot the distribution of these parsing errors for various reordering qualities. In Section 6.3.2, we illustrate these parsing errors with examples.

6.2 Preliminary Experiment

6.2.1 Gold Data

Human annotated sentences from Chinese Penn Treebank ver. 7.0 (CTB-7) are considered as the source of building up gold data. CTB-7 is a widely used corpus that comprises parsed sentences in Chinese from five genres: broadcast news (BN), broadcast conversations (BC), news magazine (NM), newswire (NS), and newsgroup weblogs (NW). We divide the corpus following the same way introduced in [106] and use the development set to obtain the gold data.

We first randomly sampled 517 unique sentences (hereinafter set-1) from all five genres in the development set. However, we note that sentences in BC and NW were mainly collected from spoken language, which tend to include faults like repetitions, incomplete sentences, corrections, or incorrect sentence segmentation. Therefore, we randomly selected another 2,126 unique sentences (hereinafter set-2) from the development set within a limit to three genres: NS, NM, and BN. Table 6.1 shows the statistics of all selected sentences in five genres respectively.

TABLE 6.1: Statistics of selected sentences in five genres of CTB-7. AL. stands for the average length of sentences, while Voc. for vocabulary. Broadcast news (BN), Broadcast conversation (BC), News magazine (NM), Newswire (NS), Newsgroups weblogs (NW).

	BN	BC	NM	NS	NW	Total
set-1	100	100	100	117	100	517
set-2	797	-	578	751	-	2,126
Total	897	100	678	868	100	2,643
AL.	29.8	20.0	33.5	28.4	25.9	29.8
Voc.	5.5K	690	5K	5.1K	972	9.5K

Meanwhile, professional human translators translated all Chinese sentences in both set-1 and set-2 into Japanese. Thereafter, according to the Japanese references, Chinese sentences in set-1 have been manually reordered to resemble the word order of their Japanese counterparts by a bilingual speaker of Chinese and Japanese for the experiments in scenario 1. For example, the Chinese sentence in Figure 2.8 is following the word order of "我(I) 东京(Tokyo) 和(and) 京都(Kyoto) 去(go to) 了(-ed)." in the handcrafted reordered set since it mirrors the Japanese word order.

After constructing these two data sets, it is necessary to convert CTB-7 parsed text to HPSG Gold-Trees and dependency Gold-Trees. HPSG parser *Chinese Enju* was used for the former one and an open utility Penn2Malt[1] [107], which converts Penn trees into MaltTab format containing dependency information, was used for the later one. Since the head rules that Penn2Malt recommends for converting on its website do not contain three new annotation types in CTB-7, we add three new ones for them as follows: FLR (Fillers) and DFL (Disfluency) head on right-hand branch; INC (Incomplete sentences) follows the same head rule as FRAG (Fragment).

6.2.2 Evaluation

To evaluate the effects of parsing errors on pre-reordering performance, we envisage two scenarios. For each scenario, we build up a benchmark for comparison and use Kendall's

[1]http://stp.lingfil.uu.se/ nivre/research/Penn2Malt.html

tau (τ) rank correlation coefficient [108] to measure the word order similaritiesin sentence pairs consisting of the benchmark data and the automatically reordered data. We use Equation 6.1 introduced in [1] to calculate the value of Kendall's tau.

$$\tau = \frac{\#\text{of concordant pairs}}{\#\text{of all pairs}} \times 2 - 1 \tag{6.1}$$

In the first scenario, we use the set of manually reordered Chinese sentences of set-1 as the benchmark and compare it with sets of automatically reordered Chinese sentences. A sentence pair example is as follows:

Manually reordered Chinese: 我(I) 东京(Tokyo) 和(and) 京都(Kyoto) 去(go to) 。

Automatically reordered Chinese: 我(I) 东京(Tokyo) 去(go to) 和(and) 京都(Kyoto) 。

Comparing with the Manually reordered Chinese, the word order of automatically reordered Chinese is "1 2 5 3 4 6", where the total number of position pairs is $\binom{6}{2}$. Pairs of the form "1 2", "1 5", or "3 6" are concordant pairs, since the value of the first position is lower than the second one. Pairs of the form "5 3" or "5 4" are not concordant, since the first position is greater than the second one. Therefore, the τ value of this Chinese sentence is $13/\binom{6}{2} \times 2 - 1 \approx 0.73$.

In the second scenario, we merge set-1 and set-2 to obtain a larger data set and the set of Japanese references plays the role of benchmark. We again compare the benchmark with sets of automatically reordered Chinese sentences generated the same way as in the first scenario. Word alignments between Chinese and Japanese are produced by MGIZA++ [97] in a file named *ch-ja.A3.final*. In this file, parallel sentence pairs (Chinese and Japanese) are aligned to each other as follows:

Chinese: 我(I) 去(go to) 东京(Tokyo) 和(and) 京都(Kyoto) 。

Japanese: NULL ( ) 私 (1) は ( ) 東京 (3) と (4) 横浜 (5) へ ( ) 行く (2) 。 (6)

The alignment order in the example is "1 3 4 5 2 6". Similarly, according to Equation 6.1, the τ value of this Chinese sentence is $11/\binom{6}{2} \times 2 - 1 \approx 0.47$.

In both scenarios, we carry out reordering methods of HFC and DPC, which are based on different parsing grammars (See Section 2.2). Accordingly, in total, there are four automatically reordered data sets that are produced by four reordering systems: Gold-Tree based reordering systems (i.e., Gold-HFC and Gold-DPC) and Auto-Tree based reordering systems (i.e., Auto-HFC and Auto-DPC). Auto-Trees are automatically generated by *Chinese Enju* and *Corbit*[2]. Gold trees are converted from CTB-7 parsed text which are created by human annotators. The baseline system uses unreordered Chinese sentences.

Due to the fact that the reordering methods are identical but the Auto-Trees may contain errors, we will be able to observe reordering differences directly caused by parsing errors. Additionally, these comparisons also reveal one's advantage between these two linguistically-motivated reordering methods.

Scenario 1 Although there are totally 517 sentences in set-1, 26 sentences were failed during the converting from CTB-7 parsed text to HPSG trees. For comparison, 491 available (Gold- and Auto-) HPSG trees and dependency trees are used to reorder sentences by two reordering methods. Our first observation on the effects of parsing errors to reordering performance is to examine word order similarities between manually reordered Chinese sentences and automatically reordered Chinese sentences. Figure 6.1 and Table 6.2 show the distribution of τ values of the 491 sentences in terms of percentage and number of sentences, respectively. Comparing to baseline, both Auto-Tree based and Gold-Tree based systems show higher average Kendall's τ values which imply that both HFC and DPC have positively reordered the Chinese sentences and improved the word alignment. Moreover, both figures show that reordering based on Gold-Trees reduced the percentage of low Kendall's τ sentences than reordering based on Auto-Trees. However, a relatively bigger improvement on Kendall's τ value distribution of Gold-DPC shows that DPC achieves better reordering quality comparing with Gold-HFC, but is sensitive on parsing errors comparing with Auto-DPC. Since the sentence number of set-1 is limited, in order to enhance the conclusions, we increased the test data by adding set-2 for further experiments in scenario 2.

[2]Note that both of *Chinese Enju* and *Corbit* was tuned with the development set of CTB-7.

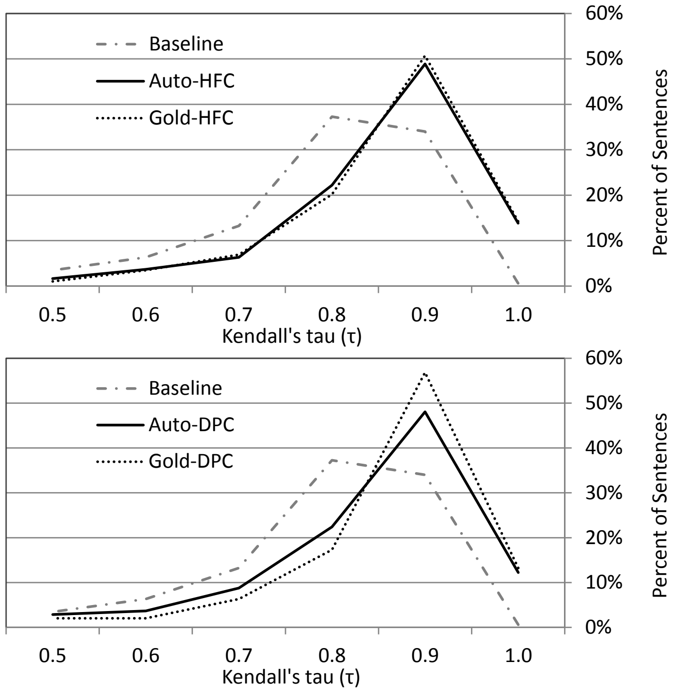


FIGURE 6.1: The distribution of Kendall's tau values for 491 sentences pairs of manually reordered with auto-reordered from the systems of baseline, Auto-HFC, Gold-HFC, Auto-DPC, and Gold-DPC.

TABLE 6.2: The distribution of Kendall's tau values for 491 bilingual sentences (Chinese-Japanese) from the systems of baseline, Auto-HFC, Gold-HFC, Auto-DPC, and Gold-DPC. (Number of sentences)

τ	Baseline	Auto-HFC	Gold-HFC	Auto-DPC	Gold-DPC
1	3	68	70	60	65
$1 \sim 0.9$	167	240	249	236	279
$0.9 \sim 0.8$	183	109	99	110	85
$0.8 \sim 0.7$	65	31	34	43	31
$0.7 \sim 0.6$	31	18	17	18	10
$0.6 \sim 0.5$	17	8	5	14	10
$0.5 \sim 0.4$	14	5	4	4	3
$0.4 \sim 0.3$	1	5	5	2	5
$0.3 \sim 0.2$	2	2	2	2	1
$0.2 \sim 0.1$	4	1	1	1	1
$0.1 \sim 0.0$	1	1	2	0	0
$-0.0 \sim -0.1$	0	0	0	0	0
$-0.1 \sim -0.2$	0	0	1	0	0
$-0.2 \sim -0.3$	0	2	0	1	1
$-0.3 \sim -0.4$	0	0	1	0	0

TABLE 6.3: Sentence numbers of monotonic alignment ($\tau == 1$) from the systems of baseline, Auto-HFC, Gold-HFC, Auto-DPC, and Gold-DPC in 491 sentence pairs. Figures with prefix of "+" are the sentence numbers that are improved comparing with baseline system, and figures with prefix of "−" are the sentence numbers that the alignments are demoted comparing with baseline system.

Baseline	Auto-HFC			Gold-HFC			Auto-DPC			Gold-DPC		
3	68	+66	−1	70	+68	−1	60	+58	−1	65	+63	−1

Scenario 2 Since we do not have manually reordered Chinese sentences as benchmark for set-2, we use Japanese references as benchmark and calculate the Kendall's tau between Chinese sentences and their Japanese counterparts by using the MGIZA++ alignment file, *ch-ja.A3.final*. The comparison implies how monotonically the Chinese sentences have been reordered to align with Japanese and there are 2,164 available (Gold- and Auto-) trees in total. Figure 6.2 shows the distribution of τ values from five systems in which the baseline is built up by using unreordered Chinese as the same way as in scenario 1.

Both figures in Figure 6.2 indicate the similar conclusions as in scenario 1, which are, 1) baseline system contains a large numbers of non-monotonic aligned sentences whereas both Gold-Trees and Auto-Trees based systems increased the amount of sentences that achieved high τ values; 2) reordering based on Gold-Trees reduced more percentage of low τ sentences; 3) specially, the amount of sentence difference in $0.9 < \tau <= 1$ between Gold-DPC and Auto-DPC shows that reordering method DPC has a high sensitivity on parsing errors; 4) furthermore, the performance of reordering system Gold-HFC and Gold-DPC sketch the figure of upper bounds of these two reordering methods. Table 6.4 shows the distribution of Kendall's tau values in terms of the number of sentences.

Discussion Tables 6.2 and 6.4 show the detailed distribution of sentences within every possible range of kendall's tau. Such distribution shows that in general, there are larger numbers of sentences that have a more similar word order to Japanese sentences (higher kendall's tau) when using HFC and DPC on gold parse trees. However, from these tables we can also observe that there are few sentences that display better kendall's tau when reordering automatically generated parse trees. Moreover, Tables 6.3 and 6.5 show that comparing with baseline system, although many sentence alignments become monotonic

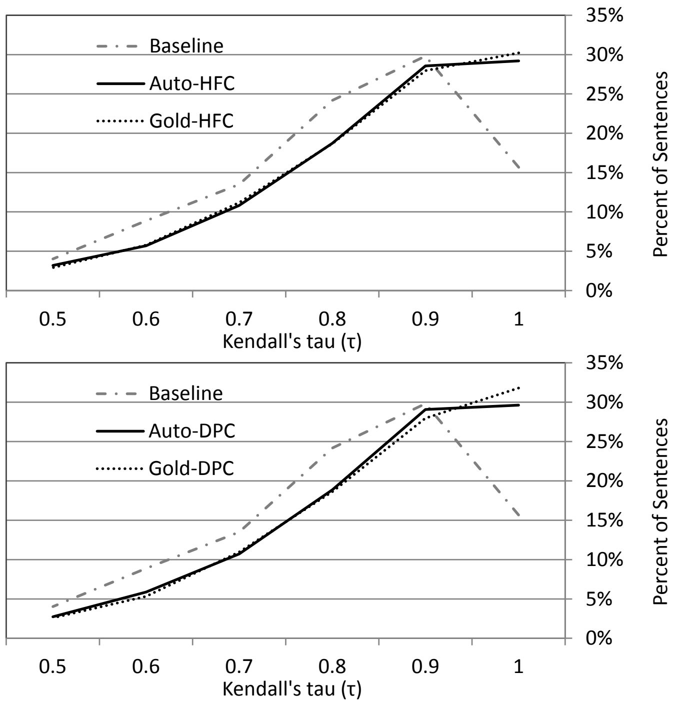


FIGURE 6.2: The distribution of Kendall's tau values for 2,164 bilingual sentences (Chinese-Japanese) from the systems of baseline, Auto-HFC, Gold-HFC, Auto-DPC, and Gold-DPC.

TABLE 6.4: The distribution of Kendall's tau values for 2,164 bilingual sentences (Chinese-Japanese) from the systems of baseline, Auto-HFC, Gold-HFC, Auto-DPC, and Gold-DPC. (Number of sentences)

τ	Baseline	Auto-HFC	Gold-HFC	Auto-DPC	Gold-DPC
1	339	632	654	641	687
$1 \sim 0.9$	645	618	605	629	608
$0.9 \sim 0.8$	523	405	404	408	403
$0.8 \sim 0.7$	292	234	242	232	236
$0.7 \sim 0.6$	192	123	125	127	114
$0.6 \sim 0.5$	87	69	63	59	55
$0.5 \sim 0.4$	42	40	37	39	35
$0.4 \sim 0.3$	11	21	14	18	15
$0.3 \sim 0.2$	16	9	8	4	4
$0.2 \sim 0.1$	6	6	5	3	3
$0.1 \sim 0.0$	4	3	1	1	1
$-0.0 \sim -0.1$	4	3	4	2	2
$-0.1 \sim -0.2$	2	1	2	1	1
$-0.2 \sim -0.3$	0	0	0	0	0
$-0.3 \sim -0.4$	1	0	0	0	0

TABLE 6.5: Sentence numbers of monotonic alignment ($\tau == 1$) from the systems of baseline, Auto-HFC, Gold-HFC, Auto-DPC, and Gold-DPC in $2,164$ sentence pairs. Figures with prefix of "+" are the sentence numbers that are improved comparing with baseline system, while figures with prefix of "−" are the sentence numbers that the alignments are demoted comparing with baseline system.

Baseline	Auto-HFC			Gold-HFC			Auto-DPC			Gold-DPC		
339	632	+414	−121	654	+424	−109	641	+414	−112	687	+457	−109

($\tau == 1$) in Gold- and Auto- reordering systems, there are monotonic sentence alignments that are demoted after reordering.

Table 6.6 lists an example in the case of HFC that we found the HFC reordering on automatically generated parse trees display a constituent order (1 and 2) more similar to the Japanese constituent order (1 and 2), when compared to the HFC reordering on gold parse trees (2 and 1). The reason for Auto-HFC to have a higher kendall's tau in this example is due to a specific translation of the Japanese sentence and the parsing err. Regarding DPC, we display in Table 6.7 a sentence that got a higher kendall's tau when reordering an automatically parsed tree using DPC when compared to reordering a gold tree. In this example, the Auto-DPC system reordered the passive particle $_1$遭(were) to the end of the sentence while the desired position is on the right-hand side of its verb $_1$泼洒(spill). Such wrong reordering should decrease the score. However, the higher kendall's tau is due to an incorrect GIZA++ alignment result, and our Kendall's taus are computed based on such auto-alignments. Therefore, we should also aware an inherent limitation of our evaluation method when using automatically aligned words.

6.3 Analysis on Cause of Reordering Errors

Preliminary experiments in Section 6.2 provide a general idea of the effects of parsing errors on reorderings. In order to achieve more explicit relationship between specific parsing errors and reordering issues, from this section, we focus on DPC and first identify concrete parsing errors by comparing dependency Gold-Trees with Auto-Trees output by *Corbit*. Since the syntactic information that guides reordering in DPC is limited to dependency structure and POS tags, for analysis on the causes of reordering errors, we

TABLE 6.6: Reordering Examples of Gold-HFC and Auto-HFC from the analysis data set. Constituents are underlined and indexed. Verbs and Particles of interest appear in red and orange, and their constituents appear in purple.

Japanese	コシュトニツァ 氏 は 、 $_1$前 国防 相 の オイダニッチ 将軍 を 含む $_2$少なくとも 14 名 高級 軍 官 を 退役 さ せる 政令 を 発表 し た 。
Original Chinese (English)	科什图尼察(Kostunica) 发布(issued) 一项(a) 政令(decree) , $_2$让(let) 至少(at least) 14名(14) 高级(senior) 军官(military officers) 退役(retire) 其中(among them) $_1$包括(including) 前(former) 国防 部长(Defense Minister) 奥伊达尼奇(Ojdanic) 将军(General) 。
Gold-HFC	科什图尼察 一 项 政令 发布 , $_2$退役 至少 14 名 高级 军官 让 其中 $_1$前 国防 部长 奥伊达尼奇 将军 包括 。
Auto-HFC	科什图尼察 一 项 政令 发布 , 其中 $_1$前 国防 部长 奥伊达尼奇 将军 包括 $_2$退役 至少 14 名 高级 军官 让 。

TABLE 6.7: Reordering Examples of Gold-DPC and Auto-DPC from the analysis data set. Verb and Particle of interest appear in red and orange.

Japanese	台北 捷 運 が 本日 午前 、 受傷 者 の 緊急 救助 演習 を 行っ た 。 15 名 の 乗客 が 不審 者 に 不明 な 液体 を $_1$撒か れ 、 重度 の やけど を 負っ た という 設定 で の 医療 緊急 救助 訓練 で ある 。
Original Chinese (English)	台北(Taipei) 捷运(MRT) 上午(morning) 进行(carry out) 了(-ed) 一场(a) 伤患(casualty) 抢救(rescue) 的 模拟(simulation) 演习(exercise) , 一共(total) 15名(15) 乘客(passengers) 模拟(were simulated) $_1$遭(were) 歹徒(criminal) $_1$泼洒(spill) 不明(unknown) 液体(liquid) 后(after) 遭到(suffered) 严重(severe) 灼伤(burns) , 而(and) 捷运(MRT) 公司(company) 立刻(immediately) 进行(co-nduct) 了(-ed) 医疗网(medical network) 的 紧急(emergency) 抢救(rescue) 。
Gold-DPC	台北 捷运 上午 一 场 伤患 抢救 的 模拟 演习 进行 了 , 一共 15 名 乘客 模拟 $_1$遭 歹徒 不明 液体 $_1$泼洒 后 严重 灼伤 遭到 , 而 捷运 公司 医疗网 的 紧急 抢救 立刻 进行 了 。
Auto-DPC	台北 捷运 上午 一 场 伤患 抢救 的 模拟 演习 进行 了 , 一共 15 名 乘客 歹徒 不明 液体 $_1$泼洒 后 严重 灼伤 遭到 , 而 捷运 公司 医疗网 的 紧急 抢救 立刻 进行 了 $_1$遭 模拟 。

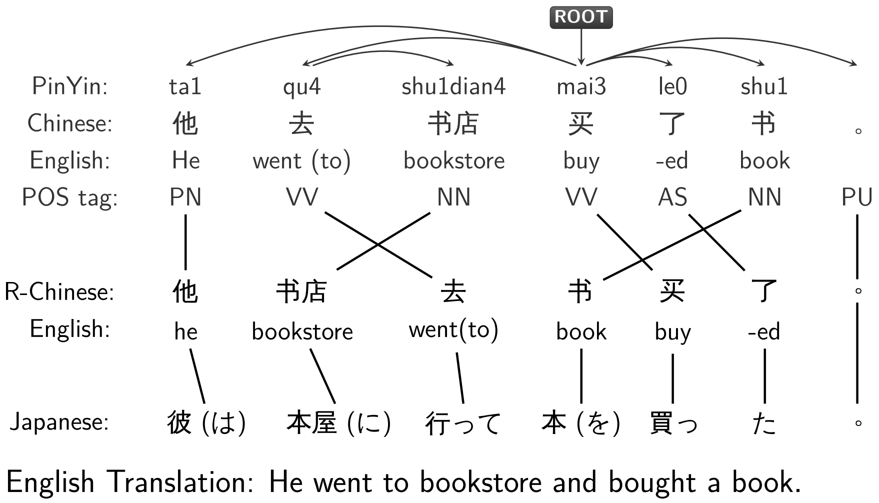
examine parsing errors from these two linguistic categories. In this section, the value of Kendall's tau measures the word order similarity between Gold-DPC and Auto-DPC.

6.3.1 Dependency parse errors by part-of-speech

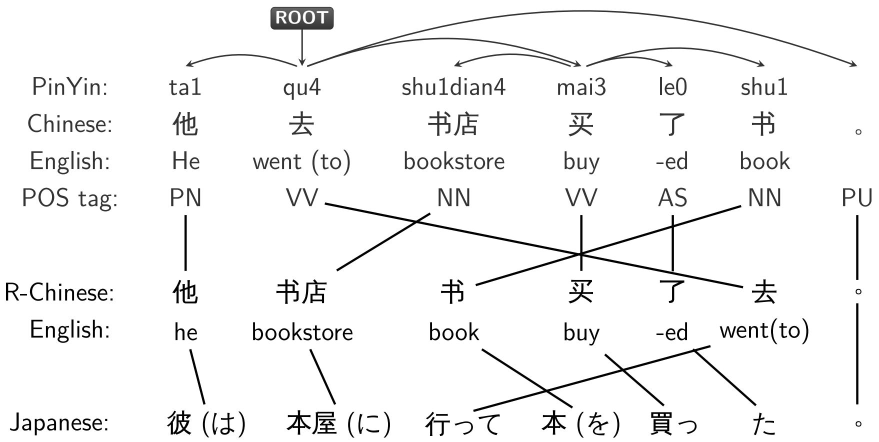
There are two types of parsing errors to a token in a dependency parse tree. One is that the token points to a wrong head, namely *dependent-error*, and another one is that the token is recognized wrongly as a head of other tokens, namely *head-error*. For example, Figure 6.3b presents a possible wrong parse tree of the example in Figure 6.3a. By comparing with the Gold-Tree in Figure 6.3b, tokens(English, POS tag) of "他(he, PN)", "去(went, VV)", "书店(bookstore, NN)", "买(buy, VV)", and ".(., PU)" in the dependency tree of Figure 6.3b all point to different wrong heads, which are dependent-errors. Concurrently, tokens(English, POS tag) of "去(went, VV)", "买(buy, VV)" are wrongly recognized as heads of other tokens (e.g., "他(he)", "书店(bookstore)"), which are head-errors. According to the definition, every head-error has at least one corresponding dependent-error. However, in the case that a token is not the root in a Gold-Tree but is root in the wrong tree, this token is a dependent-error corresponding with no head-error. As an illustration, "去(went, VV)" is the dependent-error example in Figure 6.3b.

We count the number of POS tag mis-recognitions separately for dependent- and head-errors. In the example of Figure 6.3, dependent-error counts are for VV, 2 errors, and PN, NN, PU each 1 error. The number of POS tag mis-recognitions for head-errors are VV with 2 errors. In our analysis, we will compute these counts for all POS tags at every sentence in our data set. However, our reordering method performed differently at each sentence in our data set, and the reordering quality varied from sentence to sentence. With the objective of observing the correlation between reordering quality and each type of error, we will first group sentences according to their Kendall's τ values. Then, we will compute proportions of POS tag errors at each τ value, for every type of POS tag error.

Figure 6.4 shows the distribution of top three dependent-error POS tags, which means that they are the three most frequent POS tags that point to a wrong head in auto-parse trees. VV represents all verbs except predicative adjective (VA), copula (VC), and



(A) Gold-Tree



(B) A possible erroneous parse tree

Figure 6.3: Example for calculating parsing errors in terms of POS tag.

有(have)[3] as the main verb (VE). PU represents punctuation and NN represents all nouns except proper noun (NR), temporal noun (NT), and the ones for locations which cannot modify verb phrases with or without 地(DEV)[4]. The dependent-error on VV accounts for a larger proportion in low reordering accuracy sentences whereas more NN dependent-error occurred in high reordering accuracy sentences. On the other hand, the proportion of PU dependent-error is more consistent.

Figure 6.5 shows the distribution of top two head-error POS tags, which means that they are the two most frequent POS tags that are recognized wrongly as heads in Auto-Trees. Comparing to Figure 6.4, the tendency of both VV and NN is the same but distincter.

[3] A Chinese character expresses possession and existence.
[4] A Chinese character is specially used to connect the verb phrase and its modifier.

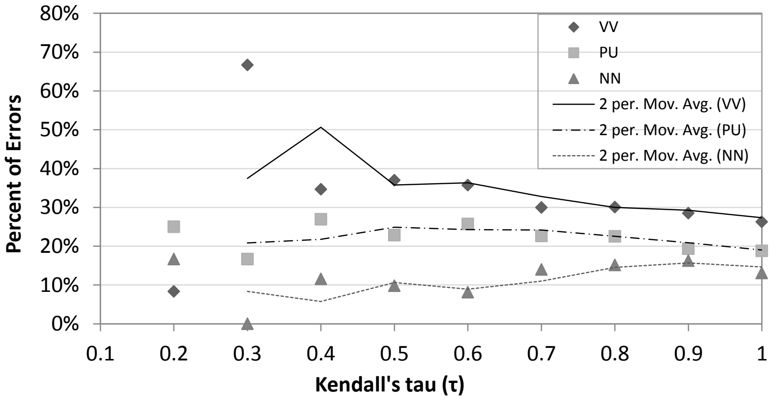


FIGURE 6.4: The distribution of top three dependent-error POS tags and their tendency lines.

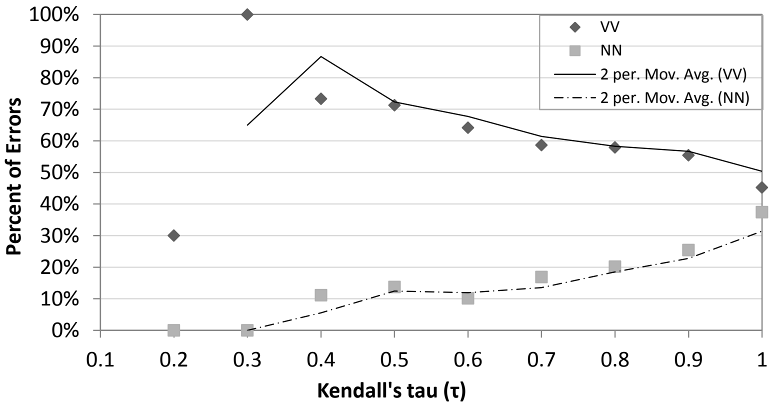


FIGURE 6.5: The distribution of top two head-error POS tags and their tendency lines.

The analysis results on the proportion distributions of dependent-error POS tags and head-error POS tags in different reordering quality sentence groups exhibit that there are more parsing errors on verbs than nouns in low reordering accuracy sentences and thus the parsing errors on verbs influence more on the reordering performance. However, it is still difficult to reveal the effects of more concrete parsing errors on reordering considering that not all verb parsing errors influence the reordering. As an illustration, in Figure 6.3, if the head of "书店(bookstore)" were "去(went)", the VV head-error of "去(went)" would not cause any reordering error since it would be reordered consistently to the right-hand side of its RM-D "书店(bookstore)". Consequently, we use a descriptive approach to analyze dependency types to explore the effects from more concrete parsing errors in the next section.

6.3.2 Dependency parse errors by dependency structure

As introduced in Section 5.1, DPC first identifies verbal block (Vb), the right-most object dependent (RM-D), and then reorders necessary words. Thus, DPC reorders not only Vb-H, but also Vb-D in a Vb, which means that the failure on identifying Vbs may also cause unexpected reordering on particles, such as aspect markers. However, in this analysis, we only focus on reordering issues of Vb-H candidates. We call it as Vb-H candidate for the reason that if it is involved into a bei-construction, then it is not Vb-H according to the conditions for being a Vb-H (See Section 5.1.1).

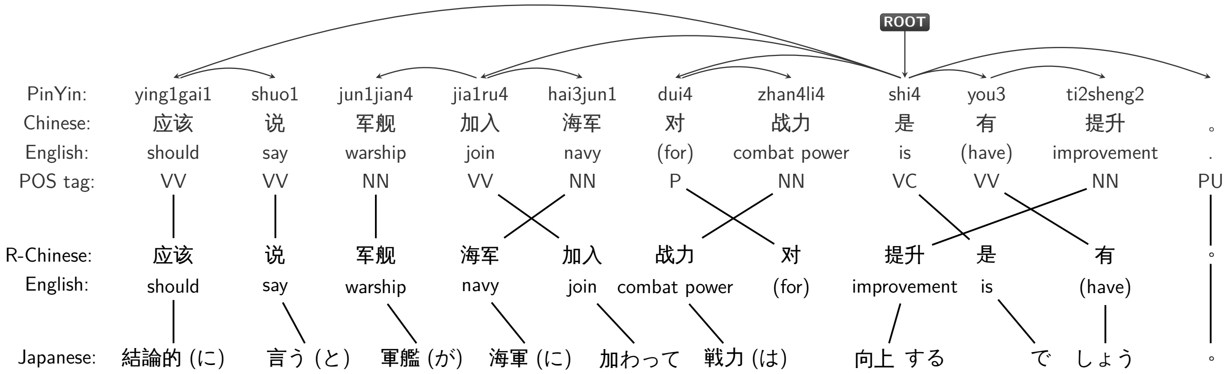
To discover the effects of more concrete parsing errors on reordering, we distinguish three categories of dependency types, namely, *ROOT*, *RM-D*, and *BEI*. Among them, ROOT denotes whether the Vb-H candidate is the root of the sentence or not, RM-D is the right-most object dependent of the Vb-H candidate if it has one, and BEI denotes whether the Vb-H candidate is involved in a bei-construction.

According to the methodology of the reordering method DPC, we define seven patterns of parsing error phenomena and classify them into three types by comparing the Gold-Tree with Auto-Tree. Table 6.8 lists all parsing error patterns in three error types, ROOT error, RM-D error and BEI error by considering three dependency types ROOT, RM-D and BEI. Symbols of "$\surd$", "×", "?" represent the status of a certain dependency type in Gold-Tree or Auto-Tree. For every Vb-H candidate, the 6 status are conditions to match the error pattern. For example, to match a Root-C error pattern, the Vb-H candidate needs to satisfy the following conditions: in Gold-Tree, it is not the root, and does not have any RM-D or bei dependent, whereas in Auto-Tree, it does not have any RM-D or bei dependent, but it is the root.

Root-A is the case where a Vb-H candidate has been wrongly parsed as the root of the sentence. However, it only affects the reordering with two constrains, namely that RM-D of the Vb-H candidate does not exist and Vb-H is not involved in a bei-construction. For instance, the Vb-H "应该(should)" in the example of Figure 6.6 was recognized as root in the Auto-Tree in Figure 6.6b. However, the actual root is the Vb-H "是(is)" in Gold-Tree of Figure 6.6a. Therefore, since "应该(should)" does not have any dependent as either

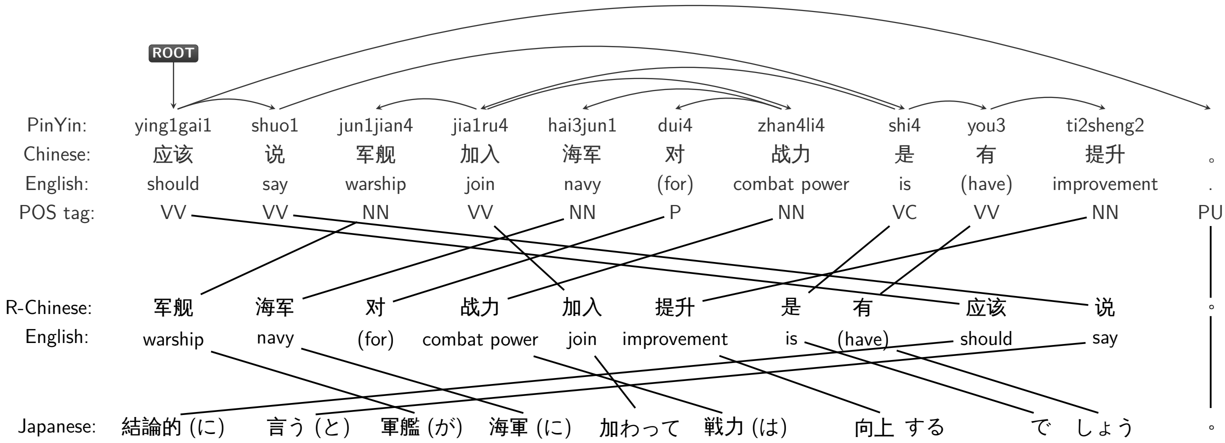
Table 6.8: Seven error patterns (Root-A, Root-G, RM_D-A, RM_D-G, RM_D-D, BEI-A, BEI-G) that cause three types of reordering issues (ROOT error, RM-D error, and BEI error). Symbols "√", "×", "?" represent the status of True, False, and Unknown, respectively. "diff." means that the RM-Ds exist in both Gold-Tree and Auto-Tree but are different.

	BEI		ROOT		RM-D	
	Gold-Tree	Auto-Tree	Gold-Tree	Auto-Tree	Gold-Tree	Auto-Tree
ROOT Error						
Root-A	×	×	×	√	×	×
Root-G	×	×	√	×	×	×
RM-D Error						
RM_D-A	×	×	×	×	×	√
	×	×	×	√	×	√
	×	×	√	×	×	√
	×	×	√	√	×	√
RM_D-G	×	×	×	×	√	×
	×	×	×	√	√	×
	×	×	√	×	√	×
	×	×	√	√	√	×
RM_D-D	×	×	×	×	√	diff.
	×	×	×	√	√	diff.
	×	×	√	×	√	diff.
	×	×	√	√	√	diff.
BEI Error						
BEI-A	×	√	√	?	×	?
	×	√	×	?	√	?
	×	√	√	?	√	?
BEI-G	√	×	?	√	?	×
	√	×	?	×	?	√
	√	×	?	√	?	√



English Translation: We can say that it improved the combat effectiveness after the warships joined the navy.

(A) Gold-Tree



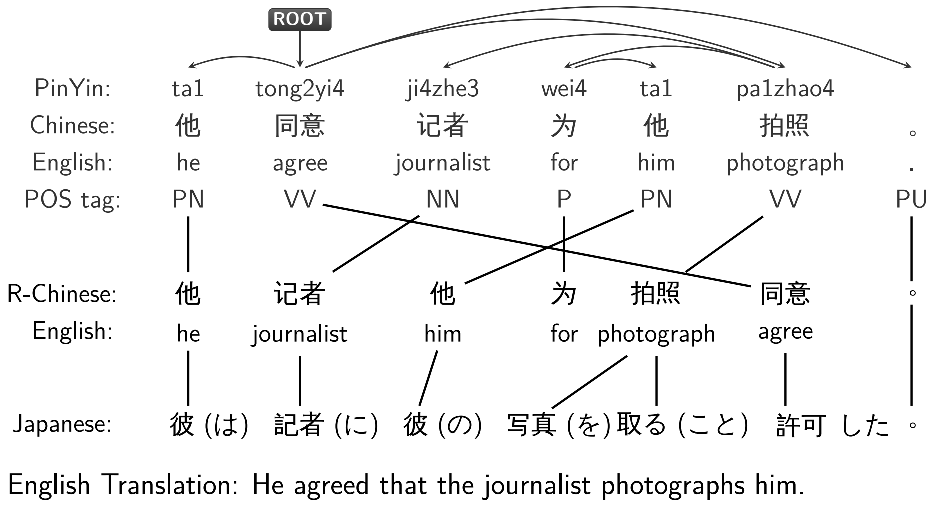
(B) A possible erroneous Auto-Tree

FIGURE 6.6: Example for parsing error patterns of Root-A and RM_D-D.

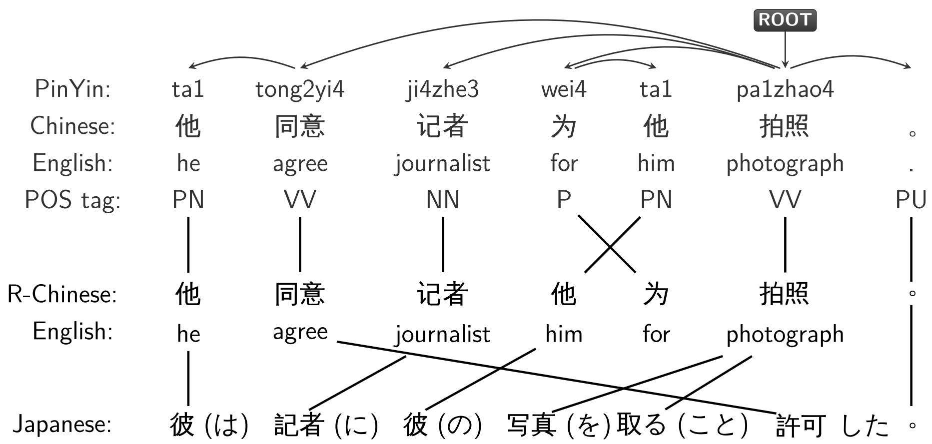
BEI or RM-D in both Gold-Tree and Auto-Tree, it will be reordered incorrectly to the end of the sentence according to the Auto-Tree whereas it will not be reordered according to Gold-Tree, which is already in the same position as its Japanese counterpart.

Root-G is the opposite case of Root-A where a Vb-H candidate is the root of the sentence but was not parsed as the root in Auto-Tree. This affects the reordering under the two same constraints as Root-A. Figure 6.7b shows an example of Root-G. In Figure 6.7a, the word alignment shows that the Vb-H "同意(agree)" should be reordered to the end of the sentence. However, it will not be reordered for the wrong parse tree shown in Figure 6.7b.

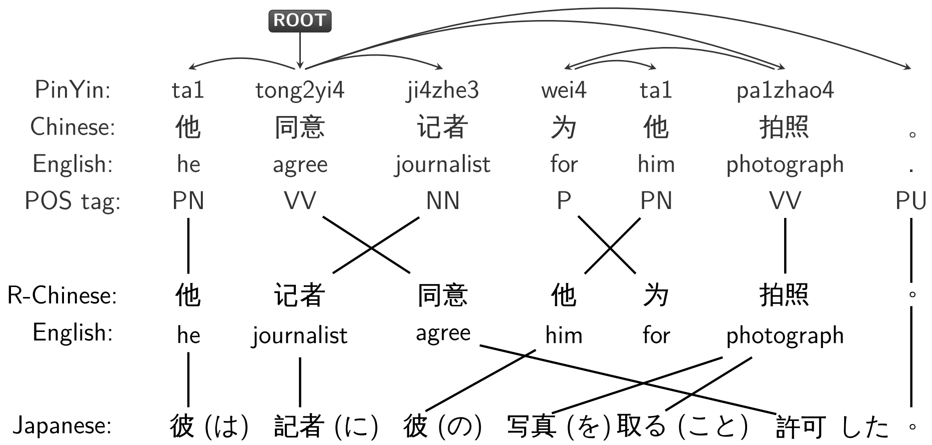
RM_D-C is the case where the RM-D of a Vb-H candidate exists in a Auto-Tree but not in Gold-Tree. In other words, a RM-D candidate was parsed wrongly on its head. There are four varieties of combination with the status of ROOT, BEI of the Vb-H candidate



(A) Gold-Tree



(B) A possible erroneous Auto-Tree



(C) Another possible erroneous Auto-Tree

Figure 6.7: Example for parsing error patterns of Root-G and RM_D-C.

that lead to incorrect reorderings. The Vb-H “同意(agree)” in Figure 6.7c matches the last combination of RM_D-C, which will be reordered right after “记者(journalist)” instead of at the end of the sentence.

RM_D-G is the opposite case of RM_D-C where the RM-D of a Vb-H candidate was missed in a Auto-Tree. There are also four cases of reordering errors according to the status of BEI, ROOT and RM-D. Vb-H “去(went to)” in Figure 6.3 matches the second combination of RM_D-G so that it will not be able to reorder after “书店(bookstore)”.

RM_D-D is the case where a bei-construction-free Vb-H candidate obtains two different RM-D candidates in Auto-Tree and Gold-Tree, which causes the reordering issue. In Figure 6.6, Vb-H “加入(join)” received different RM-Ds in two trees. According to the word alignment, it should be reordered next to “海军(navy)” instead of “战力(combat power)”.

BEI-C is the case where a Vb-H candidate received a wrong BEI dependent in Auto-Tree. This will prevent reordering independently on whether the Vb-H candidate has RM-D or is the root.

BEI-G is the opposite case of BEI-C, where Vb-H in Gold-Tree will not be reordered but in Auto-Tree it will.

After defining seven patterns of parsing errors and classifying them into three types, we calculate the average frequency proportions of each type in different τ value groups of sentences.

Figure 6.8 shows the distribution of the three types of parsing errors and their tendencies. In low τ value sentences, there are higher proportions of ROOT errors, and relatively lower proportions in high τ value sentences. RM-D errors follow the opposite tendency. This implies that the effects of ROOT errors on reordering are stronger than the effects from RM-D errors. The reason could be that ROOT errors cause long distance reordering failure while RM-D errors lead to more local reordering errors. Since there are very few BEI errors, it was difficult to capture their trends.

Figure 6.9 and Figure 6.10 provide the correlations between parsing error patterns and reordering accuracy. In ROOT errors types, Root-C had a larger percentage than Root-G

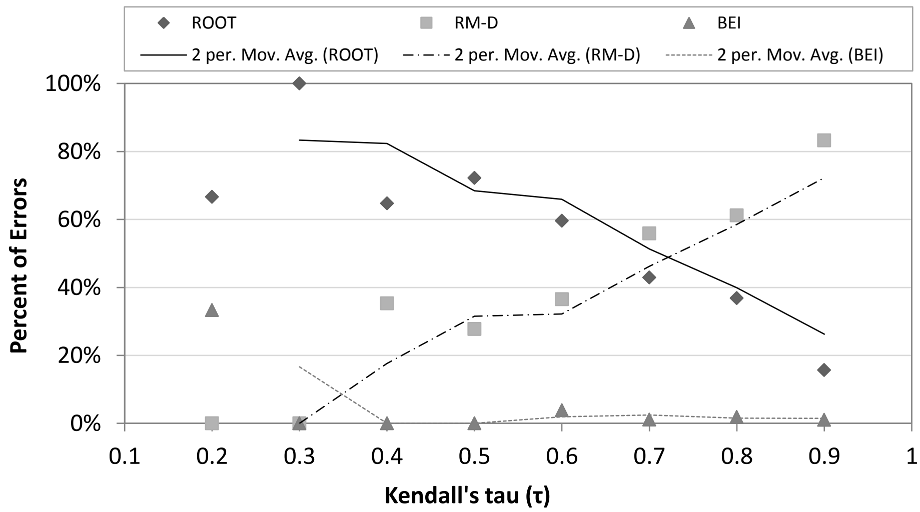


FIGURE 6.8: Distribution of three types of dependency parsing errors in different τ groups and their trend curves.

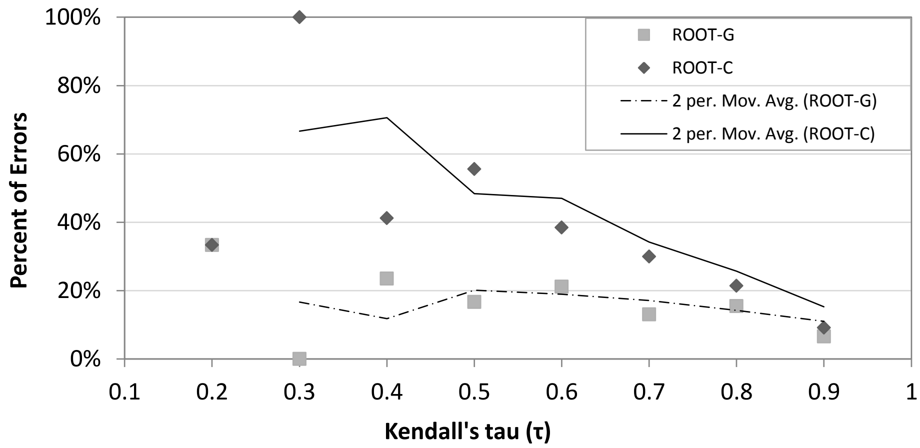


FIGURE 6.9: Distribution of patterns of ROOT error in different τ groups and their trend curves.

in low reordering accuracy sentences which shows that the Vb-H candidate that does not have any object dependent tends to be recognized as root by parser. This is consistent with the distribution results that are shown in Figure 6.10. The error pattern of RM_D-G had larger percentage than the other two patterns, which also implies that a Vb-H candidate in a Auto-Tree tends to have less or none object dependents.

6.3.3 Further Analysis Possibilities

Due to the time limitation, we only focused on analyzing parsing errors that cause reordering issues on Vb-H candidates while defining the error patterns. However, it is not

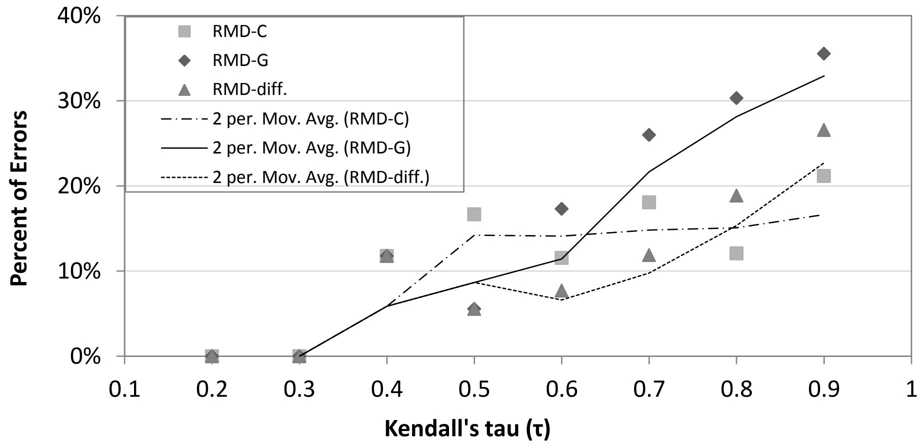


FIGURE 6.10: Distribution of patterns of RM-D error in different τ groups and their trend curves.

only that Vb-H candidates are reordered in DPC, but also other words like Vb-D candidates and particles will be reordered. It is also meaningful to explore the parsing error patterns which cause unexpected reordering on these words and the correlation between them as well. The current study on exploring influential parsing errors is not exhaustive, and another analysis possibility would be to explore what types of parsing errors do not affect reordering so that parsers can sacrifice their performance on those types of issues in order to improve on influential types.

6.4 Summary

In this chapter, we carried out linguistically motivated analysis methods by combining empirical and descriptive approaches in three analysis stages to examine the effects of different parsing errors on pre-reordering performance. We achieved four objectives: (i) quantify effects of parsing errors on reordering, (ii) estimate upper bounds in performance of the reordering method, (iii) profile general parsing errors, and (iv) examine effects of specific parsing errors on reordering.

In the first stage, we set up benchmarks in two scenarios for reordered Chinese sentences. By calculating the word order similarity between the benchmarks and the dependency parse tree based auto-reordered Chinese sentences, we quantified the correlation between parsing errors and reordering accuracies as well as explored the upper bound in reordering quality of the reordering model.

In the second stage, we examined the effects of two types of parsing errors on reordering quality by using POS tag information. The distributions of parsing errors' POS tags provide a general view of the influential parsing error types and an approximation to the cause of the effects.

In the last stage, we defined several patterns of parsing errors that assuredly cause reordering errors by using the linguistic feature of dependency types based on a deep linguistic study of the syntactic structures and the reordering model. The analysis results assist us to achieve a better and more explicit understanding on the relationship between parsing errors and reordering performance. Furthermore, we captured the effects of more concrete parsing errors on reordering.

Chapter 7

Final Remarks and Future Work

As an effective and efficient additional method in traditional SMT systems, pre-reordering starts to play an important role in the machine translation pipeline. In this chapter, we will first discuss the disadvantages of current word alignment and reordering models, and the advantages of pre-reordering methods. Then, we will discuss our HFC and DPC pre-reordering methods from both strong points and shortcomings aspects. In Section 7.2, we will exhibit several possible future directions of our current work and in Section 7.3 we will conclude this chapter.

7.1 Discussion

In this thesis, we explored syntax-informed pre-reordering for Chinese; that is, we obtain syntactic structures of Chinese sentences, reorder the words to resemble the Japanese word order, and then translate the reordered sentences using a phrase-based SMT system. However, Chinese parsers have difficulties in extracting reliable syntactic information, mainly because Chinese has a loose word order and few syntactic clues such as inflection and function words.

On one hand, parsers implementing head-driven phrase structure grammars infer a detailed constituent structure, and such a rich syntactic structure can be exploited to design well informed reordering methods. We introduced a refined reordering approach, namely HFC, by importing an existing reordering method (HF) [1] that was originally designed for English. These reordering strategies are based on Head-driven phrase structure grammars (HPSG) [10], in which the reordering decisions are made based on the head of phrases. Specifically, HPSG parsers [39, 40] are used to extract the structure of sentences in the form of binary trees. However, HFC is sensitive to parsing errors, and the binary structure of the parse trees impose hard constraints in sentences with loose word order. Moreover, as we discussed in Section 4.4 of Chapter 4, reordering strategies that are derived from the HPSG theory may not perform well when the head definition is inconsistent in the language pair under study. A typical example for the language pair of Chinese and Japanese that illustrates this phenomenon is the adverb “bu4”, which is the dependent of its verb in Chinese but the head in Japanese.

On the other hand, dependency parsers are committed to the simpler task of finding dependency relations and dependency labels, which can also be useful to guide reordering. Nevertheless, reordering methods that rely on those dependency labels will also be prone to errors, specially in the case of Chinese since it has a richer set of dependency labels when compared to other languages. In order to overcome the difficulties that we have discovered so far, in Chapter 5, we presented a hybrid approach (DPC) to pre-reorder Chinese as SVO language to improve its translation to Japanese as a SOV language, where the only required syntactic information are POS tags and unlabeled dependency parse trees. This contrasts with HFC that requires phrase structures, phrase-head information and POS tags, and the work in [63] that requires dependency relations, dependency labels and POS tags.

In spite of the fact that our DPC method uses less syntactic information, it succeeds at reordering sentences with reported speech even in presence of punctuation symbols. It is worth saying that reported speech is very common in the news domain, which might be one of the reasons of the superior translation quality achieved by the DPC pre-reordering method. DPC also accounted for ordering differences in serial verb constructions, complementizers and adverbial modifiers, which would have required an increase in the complexity of the reordering logic in other methods including HFC.

To the best of our knowledge, dependency parsers are more common than HPSG parsers across languages, and DPC can potentially be applied to translate under-resourced languages into other languages with a very different sentence structure, as long as they count with dependency parsers and reliable POS taggers.

The pre-reordering strategies discussed in this thesis were developed for Chinese to Japanese, as an example of language pair with SVO-SOV sentence structure. Thus, our findings are circumscribed to the problem of translating between SVO and SOV languages, and language pairs with similar word orders (to each other) would not benefit from using this kind of strategies. HFC and DPC reordering strategies could be applied to translating Chinese into other target languages with SOV structure, such as Chinese-Korean or Chinese-Turkish. However, these pre-reordering strategies could not be applied directly to other SVO-SOV language pairs where Chinese is not the source language, since the set of POS tags of the source languages may differ. In spite of it, we expect that dependency relations and many general POS tags hold the same properties across source languages, and we expect many reordering rules that were developed in the context of this thesis could be useful (or at least inspire) reordering rules in other SVO languages that need to be translated into an SOV language. As an example, if we were to translate English to Japanese, the rule of moving words with POS tag "VV" (verbs) to the right would still be valid for English to Japanese translation, in a similar way it was valid for Chinese to Japanese. In general, implementing DPC for other languages would first require a linguistic study on the word order differences between the two particular distant language pairs. However, some word ordering differences might be consistent across SVO and SOV language pairs (such as verbs going before or after their objects), but other ordering differences may need special treatment for the language pair under consideration (i.e. Chinese “bei4” particles).

In our evaluations, we used single-reference test sets, where Chinese sentences only had one corresponding gold Japanese translation. Unfortunately, single-reference test sets are more the rule than the exception in machine translation, due to the high cost of producing multi-reference test sets (several translations for every source sentence). In single-reference evaluations, machine translation systems may produce adequate and fluent translations that do not perfectly match the single reference, and evaluation metrics that enforce matchings of word sequences may underestimate (in absolute terms) system performance. Evaluation metrics used in state-of-the-art systems do not rely on perfect matchings of translated sentences to single-reference sentences; instead, they account for word overlaps between sequences of varying length (as in the case of BLEU), or in the relative word order between words in the system translation and the single reference (as in the case of RIBES). Automatic evaluation strategies in machine translation are subject to active discussion, as automatic metrics of translation quality define the objective functions that statistical systems attempt to optimize. Despite of this controversy, when the purpose is to compare different systems, single-reference test sets may suffice, if test sets are large enough (as we believe it was the case here) and the systems under consideration follow a similar paradigm (as in phrase-based systems in this thesis).

The results presented in this thesis report substantial differences in performance when translating sentences from news domain or from patent domain. From a human translation perspective, translating sentences from news domain should be an easier task, as sentences are generally shorter (see Table 2.1) and contain words that are more accessible to general audiences, when compared to sentences from the patent domain. Moreover, sentences from news domain display more frequently the use of reported speech, and our pre-reordering methods proved to be specially effective to handle such linguistic phenomena. However, the translation quality achieved by our systems when translating sentences from news domain was substantially lower than when translating sentences from the patent domain, which may seem counter-intuitive. We believe there are three main explanations for the differences in these results. The first one is the amount of training data that was used. In the news domain, we used around 340 and 620 thousand sentences to train the SMT systems, while in the patent domain, we used around 2.5 and 4.9 million sentences (see Table 2.1). Larger training corpus on the patent domain probably led to a higher coverage of bilingual phrases, better estimations of word-to-word alignments and better estimations of parameters in the reordering models. The second explanation would be related to the lower proportion of out-of-vocabulary words in the test and development sets of the patent domain, when compared to the sets from the news domain, which shows that our systems had a higher vocabulary coverage due to the larger training data or to a controlled vocabulary in the patent domain. The third explanation could be related to the relative regularity of syntactic structures in patent domain when compared to the news domain. Such syntactic regularity would be beneficial to the learning of reordering patterns and extraction of bilingual phrases.

Both HFC and DPC as pre-reordering strategies that use syntactic information have proved successful, but they are likely to magnify parsing errors since their reordering rules rely on parse information. This is aggravated when reordering Chinese sentences due to its loose word order and low parsing accuracy. Two important research directions concentrate on either improving parsers or developing linguistically motivated pre-reordering methods that are robust in presence of parsing errors. We believe that analyzing the link between those directions can help us to refine future developments. Accordingly, we presented a detailed analysis in Chapter 6 on observing the relationship between parsing and pre-reordering.

We found that not all POS tagging and parsing errors correlate equally with reordering quality. In the case of DPC reordering method, mis-recognitions of VV words correlate with low reordering performance, whereas mis-recognitions of NN words had a smaller impact. Indeed, DPC heavily relies on detecting verbal blocks that are candidates for reordering, and systems that use the same strategy should choose POS taggers that display high accuracy in VV recognition.

One of the key characteristics of DPC is its ability to correctly reorder sentences with reported speech constructions. For that purpose, it is crucial for parsers to recognize the sentence root, and our analysis demonstrated that systems that follow a similar strategy should rely on parsers that have a high accuracy to recognize the sentence root.

In general, we believe that future developments of syntax-based pre-reordering methods would benefit of preliminary analysis of POS tagging and parsing accuracies. In case of linguistically motivated pre-reordering methods, reordering rules could be designed to be more robust against unreliable POS tags or unreliable dependency relations. For automatically learned reordering rules, those systems could be designed to make use of N-best lists of certain POS tags or dependencies that are critical but that parsers cannot reliably provide. Additionally, researchers interested in developing POS taggers and parsers with the objective to aid pre-reordering could attempt to maximize the accuracy of POS tags or dependencies that are relevant to the reordering task, maybe at the expense of lower accuracies on other elements.

7.2 Future Work

There are two possible directions to extend the present works. The first one would be to refine the current syntax based pre-reordering methods, both HFC and DPC, to reduce its sensitivity to POS tagging or parse errors, and to extend our linguistic study on ordering differences between Chinese and Japanese languages. The second direction would be to manually or automatically find common patterns of ordering differences between SVO and SOV languages. The objective would be then to create a one-for-all reordering method that induces monotonic word alignments between sentences from distant language pairs, and that could also be easily extended to account for the unique characteristics of the source language of interest.

From the aspect of linguistics, since our work is syntax based methods and they successfully proved that syntactic information from parsers can be used to reorder words better, we also believe that using semantic information can further increase the expressive power of reordering rules. With that objective, a deep parser can be used since it provides the semantic head of nodes and can interpret sentences by using their semantic dependency. However, parsing accuracy may still be the bottleneck.

In our work of examining the relation between parsing and pre-reordering, we observed relatively small effects on reordering quality in response of parsing errors. However, reordering quality affect word alignments, which in turn affect the quality of bilingual phrases that are extracted. It would be interesting to extend the analysis work to quantify the propagation of parsing and reordering errors in SMT pipelines, to observe the factored effect on the overall MT quality. Moreover, there are other popular syntax-based pre-reordering methods that may use different types of parsing grammars, and similar analysis would also be interesting in those contexts, possibly with a larger set of gold parsed and reordered sentences.

Currently, we mainly focus our research on Chinese pre-reordering for Chinese-to-Japanese SMT. Therefore, an interesting future direction would be extend our research to other language pairs or language independent.

7.3 Conclusion

In this thesis, we consider two state-of-the-art pre-reordering methods for statistical machine translation between Chinese and Japanese languages (See Chapter 5 and Chapter 4). The first method relies on HPSG parser, and consists in swapping the head of phrases when certain conditions are met. The second method uses a dependency parser and a set of linguistically motivated reordering rules. Both methods use parsing information to guide reordering decisions, and are sensitive to parsing errors to different extents. We compare the performance of both reordering methods on the same corpus with baseline, in terms of several metrics that account for different aspects of translation quality. We proceed in Chapter 6 to analyze quantitatively and qualitatively the influence of parsing errors on these reordering methods, and profile the type of parsing errors that have the highest impact on reordering quality.

Appendix A

Summary of Part-of-Speech Tag Set in Penn Chinese Treebank

Table A.1: POS tags defined in Penn Chinese Treebank v3.0 (Xia 2000)

POS tag	Category	Instance
AD	adverb	还(yet)
AS	aspect marker	了(-ed)
BA	ba3(把) in ba-construction	把(have sth. done)
CC	coordinating conjunction	和(and)
CD	cardinal number	一百(a hundred)
CS	subordinating conjunction	虽然(although)
DEC	de0(的) in a relative-clause	的(as a complementizer or a nominalizer)
DEG	associative de0(的)	的(as a genitive marker and an associative marker)
DER	de0(得) in V-de construction and V-de-R	得(resultative)
DEV	de0(地) before VP	地(manner)
DT	determiner	这(the)
ETC	for words deng3(等), deng3deng3(等等)	等(et cetera)
FW	foreign words	ISO
IJ	interjection	啊(ah)
JJ	other noun-modifier	共同(collective)
LB	bei4(被) in long bei-construction	被(passive voice)
LC	localizer	里(inside)
M	measure word	个(piece)
MSP	other particle	所(that which)
NN	common noun	书(book)
NR	proper noun	美国(The United States)
NT	temporal noun	今天(today)
OD	ordinal number	第一(first)
ON	onomatopoeia	哈哈(ahh)
P	preposition excl. 被 and 把	从(from)
PN	pronoun	他(he)
PU	punctuation	。(.)
SB	bei4(被) in short bei-construction	被(passive voice)
SP	sentence-final particle	吗(ma)
VA	predicative adjective	红(red)
VC	shi4(是)	是(be)
VE	you3(有) as the main verb	有(have)
VV	other verb	走(walk)

Appendix B

Head Rules for Penn2Malt to Convert the Penn Chinese Treebank

Table B.1: Rules for converting trees in the Penn Chinese Treebank format into MaltTab format using Penn2Malt tool (Joakim Nivre, 2004). These rules were originally compiled by Yuan Ding, and were used to identify head branches of phrase structures. As an example, in an ADJP branch (first row), in order to discover the head branch we scan from right (r) to left all branches. If we find an ADJP or JJ branch, then we select it as a head. If we do not find them, then we scan again the branches from right (r) to left, searching for AD, NN or CS. If we do not find them, then we select the right-most (r) branch. In this work, we introduced new rules to identify head branches for FLR, INC and DFL phrases, which are not originally covered in Penn2Malt tool.

ADJP	r ADJP JJ;r AD NN CS;r
ADVP	r ADVP AD;r
CLP	r CLP M;r
CP	r DEC SP;l ADVP CS;r CP IP;r
DNP	r DNP DEG;r DEC;r
DP	l DP DT;l
DVP	r DVP DEV;r
FRAG	r VV NR NN;r
INTJ	r INTJ IJ;r
IP	r IP VP;r VV;r
LCP	r LCP LC;r
LST	l LST CD OD;l
NP	r NP NN NT NR QP;r
PP	l PP P;l
PRN	r NP IP VP NT NR NN;r
QP	r QP CLP CD OD;r
UCP	r
VCD	r VCD VV VA VC VE;r
VCP	r VCP VV VA VC VE;r
VNV	r VNV VV VA VC VE;r
VP	l VP VA VC VE VV BA LB VCD VSB VRD VNV VCP;l
VPT	r VNV VV VA VC VE;r
VRD	r VRD VV VA VC VE;r
VSB	r VSB VV VA VC VE;r
WHNP	r WHNP NP NN NT NR QP;r
WHPP	l WHPP PP P;l
FLR	r
INC	r VV NR NN;r
DFL	r

Bibliography

[1] Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh. Head finalization: A simple reordering rule for SOV languages. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and Metrics MATR*, pages 244–251. Association for Computational Linguistics, 2010.

[2] Dan Han, Katsuhito Sudoh, Xianchao Wu, Kevin Duh, Hajime Tsukada, and Masaaki Nagata. Head finalization reordering for Chinese-to-Japanese machine translation. In *Proceedings of the 6th Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-6)*, pages 57–66. Association for Computational Linguistics, 2012.

[3] Dan Han, Pascual Martínez-Gómez, Yusuke Miyao, Katsuhito Sudoh, and Masaaki Nagata. Using unlabeled dependency parsing for pre-reordering for Chinese-to-Japanese statistical machine translation. In *Proceedings of the 2nd Workshop on Hybrid Approaches to Translation (HyTra)*, pages 25–33. Association for Computational Linguistics, 2013.

[4] Dan Han, Pascual Martínez-Gómez, Yusuke Miyao, Katsuhito Sudoh, and Masaaki Nagata. Effects of parsing errors on pre-reordering performance for Chinese-to-Japanese SMT. In *Proceedings of the 27th Pacific Asia Conference on Language Information and Computing (PACLIC)*. The PACLIC Steering Committee, 2013.

[5] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[6] Franz Josef Och and Hermann Ney. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449, 2004.

[7] Richard Zens, Franz Josef Och, and Hermann Ney. Phrase-based statistical machine translation. In *Proceedings of the German Conference on Artificial Intelligence (KI 2002)*, pages 18–32. Springer, 2002.

[8] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics, 2003.

[9] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[10] Carl Jesse Pollard and Ivan A. Sag. *Head-driven phrase structure grammar*. The University of Chicago Press and CSLI Publications, 1994.

[11] Kevin Knight. Automating knowledge acquisition for machine translation. *AI Magazine*, 18(4):81, 1997.

[12] Kevin Knight. A statistical mt tutorial workbook. In *Prepared for the 1999 JHU Summer Workshop*, 1999.

[13] Adam Lopez. Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40(3):8, 2008.

[14] Philipp Koehn. *Statistical machine translation*. Cambridge University Press, 2009.

[15] Warren Weaver. Translation. *Machine Translation of Languages*, 14:15–23, 1955 (Reprinted).

[16] Bonnie J. Dorr, Pamela W. Jordan, and John W. Benoit. A survey of current paradigms in machine translation. *Advances in Computers*, 49:1–68, 1999.

[17] John Hutchins. Machine translation: A concise history. *Computer Aided Translation: Theory and Practice*, 2007.

[18] John Hutchins. Alpac: the (in) famous report. *Readings in Machine Translation*, 14:131–135, 2003.

[19] Lisette Appelo. A compositional approach to the translation of temporal expressions in the Rosetta system. In *Proceedings of the 11th Coference on Computational Linguistics*, pages 313–318. Association for Computational Linguistics, 1986.

[20] Doug Arnold and Louis Des Tombe. Basic theory and methodology in eurotra. *Machine Translation: Theoretical and Methodological Issues*, pages 114–135, 1987.

[21] Doug Arnold and Steven Krauwer. "Relaxed" compositionality in machine translation. In *Proceedings of the 2nd International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages. Session 3: EUROTRA Perspectives*, 1988.

[22] M.T. Rosetta. *Compositional Translation*. Kluwer Academic Publishers, Dordrecht, 1994.

[23] Bernard Vauquois. A survey of formal grammars and algorithms for recognition and transformation in machine translation. *IFIP Congress'68*, pages 254–260, 1968.

[24] Makoto Nagao. A framework of a mechanical translation between japanese and english by analogy principle. *Artificial and Human Intelligence (A. Elithorn and R. Banerji, editors)*, pages 173–180, 1984.

[25] Satoshi Sato and Makoto Nagao. Toward memory-based translation. In *Proceedings of the 13th Conference on Computational Linguistics-Volume 3*, pages 247–252. Association for Computational Linguistics, 1990.

[26] Osamu Furuse and Hitoshi Iida. An example-based method for transfer-driven machine translation. In *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 139–150, 1992.

[27] Harold Somers. Review article: Example-based machine translation. *Machine Translation*, 14(2):113–157, 1999.

[28] Michael Carl and Andy Way. *Recent advances in example-based machine translation*, volume 21. Springer, 2003.

[29] Peter Brown, John Cocke, S. Della Pietra, V. Della Pietra, Frederick Jelinek, Robert Mercer, and Paul Roossin. A statistical approach to language translation. In *Proceedings of the 12th Conference on Computational Linguistics-Volume 1*, pages 71–76. Association for Computational Linguistics, 1988.

[30] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990.

[31] Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, John D Lafferty, and Robert L Mercer. Analysis, statistical transfer, and synthesis in machine translation. In *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 83–100, 1992.

[32] Kevin Knight and Philipp Koehn. What's new in statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL. Tutorial*, pages 5–5, 2003.

[33] Leonard E. Baum. An equality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. *Inequalities*, 3:1–8, 1972.

[34] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[35] Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1): 39–71, 1996.

[36] Kishore A. Papineni, Salim Roukos, and Todd R. Ward. Feature-based language understanding. In *Proceedings of the 5th European Conference on Speech Communication and Technology*, pages 1435–1438, 1997.

[37] Kishore A. Papineni, Salim Roukos, and Todd R. Ward. Maximum likelihood and discriminative training of direct translation models. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 189–192. IEEE, 1998.

[38] Franz Josef Och and Hermann Ney. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 295–302. Association for Computational Linguistics, July 8–10 2002.

[39] Yusuke Miyao and Jun'ichi Tsujii. Feature forest models for probabilistic HPSG parsing. *Computational Linguistics*, 34(1):35–80, March 2008.

[40] Kun Yu, Yusuke Miyao, Takuya Matsuzaki, Xiangli Wang, and Junichi Tsujii. Analysis of the difficulties in Chinese deep parsing. In *Proceedings of the 12th International Conference on Parsing Technologies*, pages 48–57. Association for Computational Linguistics, 2011.

[41] Jun Hatori, Takuya Matsuzaki, Yusuke Miyao, and Jun'ichi Tsujii. Incremental joint POS tagging and dependency parsing in Chinese. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 1216–1224. Asian Federation of Natural Language Processing, 2011.

[42] Hong-Mei Zhao, Ya-Juan Lv, Guo-Sheng Ben, Yun Huang, and Qun Liu. *Evaluation Report for The 7th China Workshop on Machine Translation (CWMT2011)*, 2011. URL `http://mt.xmu.edu.cn/cwmt2011/document/papers/e00.pdf`.

[43] Xiaoyi Ma. Champollion: A robust parallel text sentence aligner. In *Proceedings of 5th International Conference on Language Resources and Evaluation (LREC-5)*, pages 489–492. Citeseer, 2006.

[44] Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh. HPSG-based preprocessing for English-to-Japanese translation. *ACM Transactions on Asian Language Information Processing (TALIP)*, 11(3):8:1–8:16, September 2012.

[45] Adam L. Berger, Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Andrew S. Kehler, and Robert L. Mercer. Language translation apparatus and method using context-based translation models, April 1996. United States Patent 5510981.

[46] Kevin Knight. Decoding complexity in word-replacement translation models. *Computational Linguistics*, 25(4):607–615, 1999.

[47] Christoph Tillmann and Hermann Ney. Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Computational Linguistics*, 29(1):97–133, 2003.

[48] C. Robert Moore and Chris Quirk. Faster beam-search decoding for phrasal statistical machine translation. In *Proceedings of Machine Translation Summit XI*, page 321–327. The International Association for Machine Translation (IAMT), 2007.

[49] Kenji Yamada and Kevin Knight. A syntax-based statistical translation model. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 523–530, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics.

[50] Christoph Tillmann. A unigram orientation model for statistical machine translation. In *Proceedings of the Annual Meeting of Human Language Technology Conference / North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 101–104. Association for Computational Linguistics, 2004.

[51] Michel Galley and Christopher D Manning. A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 848–856. Association for Computational Linguistics, 2008.

[52] Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of International Workshop on Spoken Language Translation (IWSLT)*, pages 68–75, 2005.

[53] Shankar Kumar and William Byrne. Local phrase reordering models for statistical machine translation. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 161–168. Association for Computational Linguistics, 2005.

[54] Kazuteru Ohashi, Kazuhide Yamamoto, Kuniko Saito, and Masaaki Nagata. Nut-ntt statistical machine translation system for IWSLT 2005. In *Proceedings of International Workshop on Spoken Language Translation (IWSLT)*, pages 128–133, 2005.

[55] Masaaki Nagata, Kuniko Saito, Kazuhide Yamamoto, and Kazuteru Ohashi. A clustered global phrase reordering model for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 713–720. Association for Computational Linguistics, 2006.

[56] Yaser Al-Onaizan and Kishore Papineni. Distortion models for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 529–536. Association for Computational Linguistics, 2006.

[57] Richard Zens and Hermann Ney. Discriminative reordering models for statistical machine translation. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 55–63. Association for Computational Linguistics, 2006.

[58] Deyi Xiong, Qun Liu, and Shouxun Lin. Maximum entropy based phrase reordering model for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 521–528. Association for Computational Linguistics, 2006.

[59] Dennis N. Mehay and Chris Brew. CCG syntactic reordering models for phrase-based machine translation. In *Proceedings of the 7th Workshop on Statistical Machine Translation*, pages 210–221. Association for Computational Linguistics, 2012.

[60] Fei Xia and Michael McCord. Improving a statistical MT system with automatically learned rewrite patterns. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, pages 508–514. Association for Computational Linguistics, 2004.

[61] Michael Collins, Philipp Koehn, and Ivona Kučerová. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 531–540. Association for Computational Linguistics, 2005.

[62] Chao Wang, Michael Collins, and Philipp Koehn. Chinese syntactic reordering for statistical machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 737–745. Association for Computational Linguistics, June 2007.

[63] Peng Xu, Jaeho Kang, Michael Ringgaard, and Franz Och. Using a dependency parser to improve SMT for subject-object-verb languages. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 245–253. Association for Computational Linguistics, 2009.

[64] Dmitriy Genzel. Automatically learning source-side reordering rules for large scale machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 376–384. Association for Computational Linguistics, 2010.

[65] Karthik Visweswariah, Jiri Navratil, Jeffrey Sorensen, Vijil Chenthamarakshan, and Nanda Kambhatla. Syntax based reordering with automatically derived rules for improved statistical machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 1119–1127. Association for Computational Linguistics, 2010.

[66] Ibrahim Badr, Rabih Zbib, and James Glass. Syntactic phrase reordering for English-to-Arabic statistical machine translation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 86–93. Association for Computational Linguistics, 2009.

[67] Ananthakrishnan Ramanathan, Hansraj Choudhary, Avishek Ghosh, and Pushpak Bhattacharyya. Case markers and morphology: addressing the crux of the fluency problem in English-Hindi SMT. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing*, pages 800–808. Association for Computational Linguistics, 2009.

[68] Young-Suk Lee, Bing Zhao, and Xiaoqiang Luo. Constituent reordering and syntax models for english-to-japanese statistical machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 626–634. Association for Computational Linguistics, 2010.

[69] Hua Wu and Haifeng Wang. Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21(3):165–181, 2007.

[70] Takashi Tsunakawa, Naoaki Okazaki, Xiao Liu, and Jun'ichi Tsujii. A Chinese-Japanese lexical machine translation through a pivot language. *ACM Transactions on Asian Language Information Processing (TALIP)*, 8(2):9:1–9:21, May 2009.

[71] Chi-Ho Li, Minghui Li, Dongdong Zhang, Mu Li, Ming Zhou, and Yi Guan. A probabilistic approach to syntax-based reordering for statistical machine translation. In *Proceedings of the 45rd Annual Meeting on Association for Computational Linguistics (ACL)*, volume 45, pages 720–727. Association for Computational Linguistics, 2007.

[72] Xianchao Wu, Katsuhito Sudoh, Kevin Duh, Hajime Tsukada, and Masaaki Nagata. Extracting pre-ordering rules from predicate-argument structures. In *Proceedings of 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 29–37. Asian Federation of Natural Language Processing, 2011.

[73] Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, and Christopher D Manning. Discriminative reordering with Chinese grammatical relations features. In *Proceedings of the 3rd Workshop on Syntax and Structure in Statistical Translation*, pages 51–59. Association for Computational Linguistics, 2009.

[74] Marta R Costa-Jussà and José AR Fonollosa. Statistical machine reordering. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 70–76. Association for Computational Linguistics, 2006.

[75] Kay Rottmann and Stephan Vogel. Word reordering in statistical machine translation with a pos-based distortion model. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, pages 171–180, 2007.

[76] Roy Tromble and Jason Eisner. Learning linear ordering problems for better translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 1007–1016. Association for Computational Linguistics, 2009.

[77] Karthik Visweswariah, Rajakrishnan Rajkumar, Ankur Gandhe, Ananthakrishnan Ramanathan, and Jiri Navratil. A word reordering model for improved machine translation. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 486–496. Association for Computational Linguistics, 2011.

[78] Graham Neubig, Taro Watanabe, and Shinsuke Mori. Inducing a discriminative parser to optimize machine translation reordering. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 843–853. Association for Computational Linguistics, 2012.

[79] Katsuhito Sudoh, Xianchao Wu, Kevin Duh, Hajime Tsukada, and Masaaki Nagata. Post-ordering in statistical machine translation. In *Proceedings of the 13th Machine Translation Summit*, page 316–323. The International Association for Machine Translation (IAMT), 2011.

[80] Isao Goto, Masao Utiyama, and Eiichiro Sumita. Post-ordering by parsing for japanese-english statistical machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, pages 311–316. Association for Computational Linguistics, 2012.

[81] Chris Quirk and Simon Corston-Oliver. The impact of parse quality on syntactically-informed statistical machine translation. In *Proceedings of Empirical Methods on Natural Language Processing (EMNLP)*, pages 62–69. Association for Computational Linguistics, 2006.

[82] Nathan Green. Effects of noun phrase bracketing in dependency parsing and machine translation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT), Student Session*, pages 69–74. Association for Computational Linguistics, 2011.

[83] Jason Katz-Brown, Slav Petrov, Ryan McDonald, Franz Och, David Talbot, Hiroshi Ichikawa, and Masakazu Seno. Training a parser for machine translation reordering. In *Proceedings of Empirical Methods on Natural Language Processing (EMNLP)*, pages 183–192. Association for Computational Linguistics, 2011.

[84] Tadayoshi Hara, Yusuke Miyao, and Jun'ichi Tsujii. Descriptive and empirical approaches to capturing underlying dependencies among parsing errors. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1162–1171. Association for Computational Linguistics, 2009.

[85] Ryan McDonald and Joakim Nivre. Characterizing the errors of data-driven dependency parsing models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 122–131. Association for Computational Linguistics, 2007.

[86] Mark Dredze, John Blitzer, Partha Pratim Talukdar, Kuzman Ganchev, Joao Graca, and Fernando Pereira. Frustratingly hard domain adaptation for dependency parsing. In *Proceedings of the CoNLL Shared Task Session of the 2007 Joint*

Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 1051–1055. Association for Computational Linguistics, 2007.

[87] Jesús Giménez and Lluis Màrquez. Towards heterogeneous automatic MT error analysis. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, pages 1894–1901. European Language Resources Association, 2008.

[88] Vivian James Cook and Mark Newson. *Chomsky's Universal Grammar: An introduction*. Oxford: Basil Blackwell, 1988.

[89] Naoki Fukui. *Theory of Projection in Syntax*. CSLI Publisher and Kuroshio Publisher, 1992.

[90] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas (AMTA)*, pages 223–231. The Association for Machine Translation in the Americas, 2006.

[91] Qian Gao. Word order in mandarin: Reading and speaking. In *Proceedings of the 20th North American Conference on Chinese Linguistics (NACCL-20)*, volume 2, pages 611–626, 2008.

[92] Fei Xia. *The part-of-speech tagging guidelines for the Penn Chinese Treebank 3.0*, 2000.

[93] Taku Kudo and Yuji Matsumoto. Japanese dependency structure analysis based on support vector machines. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical Methods in Natural Language Processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*, pages 18–25. Association for Computational Linguistics, 2000.

[94] Pi-Chuan Chang, Michel Galley, and Christopher D Manning. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the*

3rd Workshop on Statistical Machine Translation, pages 224–232. Association for Computational Linguistics, 2008.

[95] Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440. Association for Computational Linguistics, 2006.

[96] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics, 2007.

[97] Qin Gao and Stephan Vogel. Parallel implementations of word alignment tool. In *Proceedings of Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57. Association for Computational Linguistics, 2008.

[98] Andreas Stolcke. SRILM – an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing*, pages 901–904, September 16–20 2002.

[99] Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167. Association for Computational Linguistics, 2003.

[100] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[101] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952. Association for Computational Linguistics, 2010.

[102] Christoph Tillmann, Stephan Vogel, Hermann Ney, Alex Zubiaga, and Hassan Sawaf. Accelerated dp based search for statistical translation. In *Proceedings of the 5th European Conference on Speech Communication and Technology*, pages 2667–2670, 1997.

[103] Charles N. Li and Sandra A. Thompson. *Mandarin Chinese: A functional reference grammar*. Linguistics-Asian studies. University of California Press, 1989.

[104] Yuqing Guo. *Treebank-based acquisition of Chinese LFG resources for parsing and generation*. PhD thesis, Dublin City University, 2009.

[105] Angela Downing and Philip Locke. *English grammar: a university course*. Second edition. Prentice Hall international English language teaching. Abingdon, Oxon: Routledge, 2006.

[106] Yiou Wang, Jun'ichi Kazama, Yoshimasa Tsuruoka, Wenliang Chen, Yujie Zhang, and Kentaro Torisawa. Improving Chinese word segmentation and POS tagging with semi-supervised methods using large auto-analyzed data. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 309–317. Asian Federation of Natural Language Processing, 2011.

[107] Joakim Nivre. *Inductive dependency parsing*. Springer, 2006.

[108] Maurice G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2): 81–93, 1938.