NAIST-IS-DT0261006

Doctor's Thesis

Automatic Construction of Translation Knowledge for Corpus-based Machine Translation

Kenji Imamura

May 10, 2004

Department of Information Processing
Graduate School of Information Science
Nara Institute of Science and Technology

Doctor's Thesis
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
DOCTOR of ENGINEERING

Kenji Imamura

Thesis committee:	Yuji Matsumoto, Professor
	Shunsuke Uemura, Professor
	Kiyohiro Shikano, Professor

Automatic Construction of Translation Knowledge for Corpus-based Machine Translation[*]

Kenji Imamura

Abstract

Many machine translation (MT) systems that utilize the knowledge automatically acquired from bilingual corpora have been proposed in conjunction with efforts to accumulate corpora. We call this approach corpus-based machine translation in this thesis. This thesis focuses on automatic construction of the translation knowledge needed for corpus-based MT and discusses the following three tasks.

1. Proposing a knowledge acquisition method from bilingual corpora.
2. Applying the acquired knowledge to an actual MT engine and measuring the MT quality.
3. Identifying the inherent problems of the corpus-based MT that decrease MT quality and proposing solutions.

A feature of this thesis is not only investigating the first task but also investigating the second and third tasks. In order to clarify features of corpus-based MT, this thesis identifies inherent problems by translating sentences using acquired knowledge and proposes solutions.

For the first task, this thesis proposes a hierarchical phrase alignment (HPA) method. This method automatically extracts equivalent phrases, which are corresponding expressions between bilingual sentences. HPA employs parsers. Previous methods extract correspondences after determining the parsing trees of the bilingual sentence, while HPA simultaneously extracts the best parsing trees and corresponding phrases by utilizing the structural similarity measure called the phrase correspondence score. HPA has two features. One is the ability to resolve

[*] Doctor's Thesis, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DT0261006, May 10, 2004.

parsing ambiguities by using similarity. The other feature is the ability of HPA to output the sequence of partial trees even if the parsing has failed. Using this method, about twice as many equivalent phrases were extracted than the previous methods, and almost no deterioration was observed (Chapter 2).

For the second task, HPA is applied to a large corpus, and the translation knowledge is automatically constructed. The knowledge is integrated into the MT engine, which is based on transfer driven machine translation (TDMT). Then, translation quality is measured. Through this integration, the problem of the knowledge containing many redundant rules becomes clear. These cause incorrect MT results or increase ambiguity. By eliminating the redundant rules, the MT quality of the system constructed from bilingual corpora become close to the hand-coded one (Chapter 3).

For the third task, this thesis provides two approaches. One is preprocessing in the knowledge acquisition stage while focusing on the problems caused by the bilingual corpora themselves. Not all sentences in corpora are appropriate for MT. For example, bilingual corpora usually contain context/situation dependent translation or multiple translations even though the source sentences are equal. If we constructed the knowledge from such corpora, many redundant rules would be generated. This thesis first discusses what kind of bilingual sentences are appropriate for MT and then focuses on literalness. A translation correspondence rate is defined to measure the literalness. Two knowledge construction methods are proposed. One is to filter the corpus, which collects high literal bilingual sentences before knowledge acquisition. This method could not dramatically improve the MT quality because it removed the necessary translations for MT, such as idiomatic expressions. The other is the split-construction method, which divides a bilingual sentence into literal parts and the other parts before different generalizations are applied. By using the split construction, the MT quality was improved by about 8.6% (Chapter 4).

The other approach to the third task is post-processing in the knowledge acquisition stage. Redundant rules are created not only by translation variety but also by acquisition errors. To overcome this problem, this thesis proposes the feedback cleaning method, which removes redundant rules based on the automatic evaluation of MT quality. This method regards the removal of such rules as a combinatorial optimization problem. Specifically, automatic evaluation is regarded as an objective function of the optimization, and the method searches for the optimal combination as a way to maximize the evaluation scores. Here, BLEU is used for the automatic evaluation. The hill-climbing algorithm, which involves features of this task, is applied to the process of searching for the optimal combination of rules. However, this method requires a specific evaluation corpus. To avoid this problem, we propose the N-fold cross-cleaning method, which uses the training corpus as the evaluation corpus. Cross-cleaning could considerably improve MT quality compared with the previous methods (Chapter 5).

Finally, this thesis introduces current topics and discusses future directions in corpus-based MT.

Keywords:

machine translation (MT), bilingual corpora, example-based MT, automatic knowledge construction, hierarchical phrase alignment, translation quality, controlled translation, literalness, feedback cleaning.

コーパスベース機械翻訳における翻訳知識自動構築法の研究*

今村 賢治

内容梗概

対訳コーパスの充実に伴い、機械翻訳も徐々にコーパスから自動獲得した知識を用いる方式が盛んになっている。これを本論文ではコーパスベース機械翻訳と呼ぶ。本論文では、コーパスベース機械翻訳のうち、対訳コーパスからの機械翻訳知識自動構築法に焦点を当て、以下の 3 点の課題について論じる。

1. 対訳コーパスからの知識獲得法の提案
2. 自動獲得された知識の機械翻訳エンジンへの適用と翻訳品質の測定
3. 翻訳品質を低下させるコーパスベース機械翻訳特有の問題点の指摘と解決法の提案

本論文の特徴は、第 1 の課題のみでなく、第 2、第 3 の課題に言及したところにある。コーパスベース機械翻訳の特徴を明らかにするため、本論文では、対訳コーパスから獲得した知識を用いて実際に機械翻訳を行うことにより、その問題点を指摘し、解決方法を提案する。

まず、第 1 の課題に対し、本論文では対訳文の階層的句アライメント方法を提案する。これは、同等句と呼ぶ対訳文間の句同士の対応を自動的に抽出する方法である。本論文で提案する方式は構文解析を用いる。従来、構文解析を併用する句アライメント方式は、対訳文の構文構造を決定したのちに対応関係を抽出していた。本論文で用いた方式は、句対応スコアと呼ぶ構造類似性評価尺度を用い、アジェンダ中から最良の構文構造と対応関係を同時に決定する。その利点はまず、構文解析の曖昧性を、対訳文の類似性を利用して解消することである。もう一つの利点は、構文解析に失敗しても部分木の列を出力することである。このような方式を用いることにより、従来法に比べ、精度を保持したまま約 2 倍の同等句を抽出することができた (第 2章)。

第 2 の課題に対しては、階層的句アライメントを大規模対訳コーパスに適用し、機械翻訳知識を自動構築する。さらに、実際の機械翻訳エンジン TDMT に組み込

[*] 奈良先端科学技術大学院大学　情報科学研究科　情報処理学専攻　博士論文, NAIST-IS-DT0261006, 2004 年 5 月 10 日.

み、翻訳品質を測定する。組み込む過程で、翻訳知識中には大量の冗長規則が含まれ、誤訳や曖昧性増大の原因となることが判明した。冗長規則を排除することにより、自動構築した機械翻訳器は、Hand-coded 機械翻訳 (手作業で作成した知識を用いる機械翻訳) の性能に近づくことができるという結論を得た (第 3章)。

第 3 の課題に対しては、二つのアプローチが提案される。一つは、対訳コーパス自身に起因する問題に焦点を当て、それを知識獲得の前処理的に解決する方法である。対訳コーパスは必ずしも機械翻訳に適した対訳文ばかりではなく、翻訳の多様性に起因して、文脈や状況に依存した訳や、同じ原文であるにも関わらず、異なる複数の訳が含まれる。このようなコーパスから機械翻訳知識を自動構築すると、大量の冗長規則が作成される、本論文ではまず、機械翻訳に適した対訳とは何か、議論し、直訳性に着目する。直訳性を測定するスコアとして対訳対応率を定義し、二つの知識構築法を提案する。一つは、直訳性が高い対訳文だけを用いて知識を構築する、対訳コーパスのフィルタリング法である。この方法は、慣用表現など、低直訳であるにも関わらず翻訳に必要な対訳文を排除してしまうため、大幅な品質改善はできなかった。もう一つは、対訳文を直訳部と意訳部に分割し、そこから生成される規則に異なる一般化を適用した、分割構築である。分割構築を行うことにより、約 8.6% の文について翻訳品質を改善することができた (第 4章)。

第 3 の課題に対するもう一つのアプローチは、機械翻訳知識自動獲得の後処理による解決法である。冗長な規則は、翻訳の多様性の他に、自動獲得エラーによっても作られる。この問題に対して、翻訳品質の自動評価を利用した機械翻訳規則の取捨選択法 (フィードバッククリーニング) を提案する。これは、翻訳規則の取捨選択を組み合わせ最適化問題として捉えている。すなわち、自動評価が出力するスコアを組み合わせ最適化の目標関数値と見なし、これを最大化するように、規則の取捨選択を行う。自動評価方法には BLEU を用い、組み合わせ最適化法には、機械翻訳というタスクの特徴を活かした山登り法を用いた。この方式は、自動獲得に用いたコーパス (訓練コーパス) とは別の評価コーパスを必要とするが、訓練コーパスのみでクリーニングを行うために、さらに本論文では N 分割交差クリーニングを提案する。その結果、従来の後処理法に比べ、翻訳品質を大幅に改善することができた (第 5章)。

本論文では最後にコーパスベース機械翻訳の最新トピックを紹介し、今後の方向性について論じる。

キーワード

機械翻訳, 対訳コーパス, 用例翻訳, 知識の自動構築, 階層的句アライメント, 翻訳品質, 制限対訳, 直訳性, フィードバッククリーニング.

Acknowledgments

First of all, I would like to thank Kadokawa Publishers, who permitted us to use the coding system of Ruigo-shin-jiten for research purpose.

I would like to gratitude to Prof. Yuji Matsumoto, the advisor of my doctor's thesis. When I was absent at the laboratory due to my office work, the professor gave me a lot of advice on my research. Moreover, I am grateful to Assoc. Prof. Kentaro Inui and the members of Computational Linguistics Laboratory, who frankly discussed my research topics. I would also like to thank Prof. Shunsuke Uemura and Prof. Kiyohiro Shikano for their helpful comments.

The work reported in this thesis was mainly done at Advanced Telecommunication Research Laboratories (ATR). Therefore, I have to thank the members of ATR Spoken Language Translation Research Laboratories. Dr. Eiichiro Sumita gave me advice through my research. Mr. Satoshi Shirai, a former Head of Department 3, and Dr. Hiromi Nakaiwa, Head of Department 3, supervised my research. Dr. Seiichi Yamamoto, Director of ATR-SLT, gave me a chance to research in machine translation.

Mr. Yasuhiro Akiba, Mr. Takao Doi, Dr. Andrew Finch, Mr. Hideo Okuma, Mr. Michael Paul, Mr. Mitsuo Shimohata, Dr. Taro Watanabe, Mr. Setsuo Yamada, and Mr. Nobutaka Yoshioka cooperated with me to research new machine translation methods. Ms. Etsuko Kanazaki, Ms. Saori Mine, and Ms. Mieko Taka built bilingual corpora of English and Japanese. Ms. Noriko Koyama and Mr. David Brown evaluated machine translation results. I am grateful to all former and current members of English-to-Japanese machine translation team.

I would also like to thank Mr. Yoshifumi Ooyama, Mr. Hisashi Ohara, Mr. Masanobu Higashida, and Mr. Yasuo Sakama, Group Leaders of Nippon Telegraph and Telephone Corporation (NTT). They gave me experiences in the research of natural language processing.

The people recorded here are a part of my gratitude. Finally, I would like to thank all people that took part in my research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Along with the efforts made to enlarge corpora, quite a few natural language processing systems that utilize knowledge acquired from corpora have been developed. Machine translation (MT) is not an exception. Many MT methods that utilize the knowledge acquired from bilingual corpora have been proposed. We call the approach of such MT methods *corpus-based machine translation* in this thesis.

Corpus-based machine translation is the opposite concept of 'hand-coded machine translation' (Figure 1.1). In hand-coded MT, the translation knowledge is constructed by humans. On the contrary, the knowledge of the corpus-based MT, such as rules or models, is automatically acquired and constructed from bilingual corpora.

In this chapter, the author clarifies the objective of the thesis while contrasting corpus-based MT with hand-coded MT.

1.1 Corpus-based Machine Translation

1.1.1 Problems of Hand-coded Machine Translation

Until the early nineties, the translation knowledge used by machine translation systems was constructed manually. For example, the ALT-J/E system (Ikehara et al., 1991) contained about 15,000 pattern-pairs that represented corresponding Japanese case frames and English expressions. The pattern-pairs were constructed fully manually. ALT-J/E was a high-accuracy translation system; however, generally speaking, there were the following problems with hand-coded MT, and thus it succeeded only partially from a practical point of view.

- Huge costs are incurred for constructing knowledge. As the knowledge grows

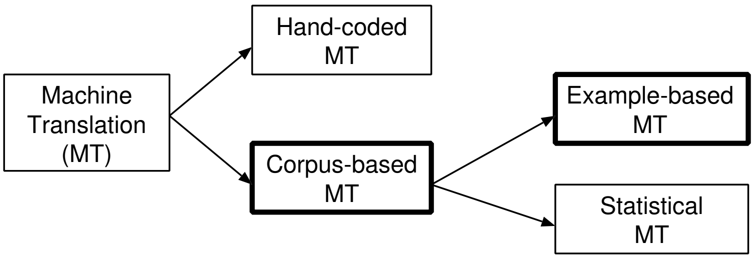


Figure 1.1. Hierarchy of Machine Translation Methods

larger, it conflicts with other existing knowledge, and the cost increases exponentially.

- The knowledge cannot be easily adapted to the new target domain. For example, if an MT system developed for a newswire was applied to the translation of technical papers, various adaptations would be necessary. Likewise, if we developed a spoken-language MT system based on a written-language MT system, we would have to adapt the knowledge manually. This work is very hard and requires long time.
- Building multilingual MT systems is difficult. If we have a Japanese-to-English MT system, we have to re-construct the knowledge in order to build an English-to-Japanese MT system.

1.1.2 Bilingual Corpora

Along with the expansion of electronic documents, many bilingual corpora have been published in many languages. For example, the Canadian Hansard Corpora (the records of the Canadian Parliament) are bilingual corpora in English and French that correspond to articles. Brown et al. (1991) obtained the articles for the period of 1973 to 1986 and then carried out the sentence alignment. Consequently, they acquired about 2.8 million bilingual sentences. The Canadian Hansard Corpora can be obtained from Linguistic Data Consortium (LDC) [1]. The version edited by Ulrich Germann is available from the Information Science Institute, University of Southern California [2].

Bilingual corpora have been developed in other languages. LDC provides other corpora such as Hong Kong Hansard (English and Chinese, about 10 million words) and UN Parallel Text (English, French, and Spanish, about 38 million

[1] http://www.ldc.upenn.edu/

[2] http://www.isi.edu/natural-language/download/hansard/index.html

words). In Japanese, Utiyama and Isahara (2003) obtained about 150,000 bilingual sentences (Japanese and English) by sentence alignment from comparable newspaper articles.

Although the above corpora contain written languages or transcriptions of a speaker reading a manuscript, corpora of naturally spoken language have also been developed. For example, Furuse et al. (1994) developed a dialogue corpus (English and Japanese, about 16,000 sentences) based on the target domain of travel conversation. The Verbmobil project (Wahlster, 2000) developed German-English dialogue corpora (about 58,000 sentences) based on the target domain of appointment negotiations and travel arrangements. In addition, Takezawa et al. (2002) and Kikui et al. (2003) are now developing a large collection of multilingual sentences (Japanese, English, Chinese, and Korean) that are usually found in phrasebooks for foreign tourists.

1.1.3 Major Approaches to Corpus-based Machine Translation

In order to solve the problems of hand-coded MT, corpus-based MT has been proposed. Corpus-based MT automatically acquires the translation knowledge or models from bilingual corpora. Currently, there are two major approaches to corpus-based MT: example-based machine translation and statistical machine translation (Figure 1.1).

Example-based Machine Translation (EBMT) Example-based machine translation originated in the translation method based on the analogy proposed by Nagao (1984). Example-based MT regards the bilingual corpus as a kind of a database. Here, an example base, which contains bilingual sentences or partial sentences of two languages, is prepared in advance. Example-based MT translates the input sentence in two steps (Somers, 1999). First, it retrieves the most similar example to the input sentence. Second, the target part of the example is modified/transformed based on the differences between the input and the example. A feature of example-based MT is its usage of thesauri. If there is no example that exactly matches the input sentence, the distance between the input and the example is measured by using the thesauri to obtain the nearest example.

Few example-based MT systems use ‘raw’ bilingual sentences as the examples. In order to increase coverage, most example-based MT systems use examples made from bilingual sentences that are decomposed or generalized. For instance, Auerswald (2000) maintains examples by using templates that generalize the date or time words as the variables. The MSR-MT system (Richardson

Table 1.1. Examples of Corpus-based Machine Translation Systems

Translation Unit	Example-based MT	Statistical MT
Word	—	(Brown et al., 1993) Most Statistical MT
Phrase	(Richardson et al., 2001) (Imamura, 2002) (Aramaki et al., 2003)	(Yamada and Knight, 2001) (Charniak et al., 2003) (Zens and Ney, 2003)
Sentence	(Auerswald, 2000) (Sumita, 2003)	—

et al., 2001) maintains partial trees of syntactic structure (called logical forms) as the examples. Since the generalized or decomposed examples are regarded as rules, we call them the *transfer rules*. Namely, automatic construction of MT knowledge for example-based MT is equivalent to the automatic decomposition and generalization of bilingual sentences.

Statistical Machine Translation (SMT) Statistical machine translation decodes the input sentence into the output sentence in the same manner of cryptanalysis (Brown et al., 1993). When the Japanese word sequence J is given, the English word sequence E, which satisfy the following equation, is searched for.

$$\operatorname*{argmax}_{E} P(E|J) = \operatorname*{argmax}_{E} P(E)P(J|E) \qquad (1.1)$$

The model represented by $P(E)$ is called the language model, and that by $P(J|E)$ is called the translation model. These models are automatically estimated from the bilingual corpora. Several methods can be used according to the model type.

Table 1.1 shows the recent corpus-based MT systems classified by their approaches and translation units. The mainstream example-based MT uses a phrase or a sentence as the translation unit, while the mainstream statistical MT uses a word.

1.1.4 Advantages of Corpus-based Machine Translation

Corpus-based machine translation has the following advantages over hand-coded machine translation.

- The cost of the knowledge construction is nearly equal to the cost of collecting bilingual corpora. Manual construction of the knowledge requires specialists who are not only familiar with both languages but who are also familiar with the machine translation architecture. However, bilingual corpora can be collected by anyone who knows both languages.

- High trans-domain portability: when we change the target domain, we only have to collect bilingual corpora of the new domain.

- It offers the potential to easily construct multilingual MT systems. At least we know that it is easy to construct Japanese-to-English and English-to-Japanese MT systems from one English-Japanese bilingual corpus.

1.2 Research Target

As described above, corpus-based machine translation has advantages in particular from the viewpoint of cost. However, the features and problems of corpus-based MT are still unclear. Corpus-based MT must have problems that are different from those of hand-coded MT. These issues include, for example, how to acquire MT knowledge, how many bilingual sentences are necessary, what type of bilingual corpora is suitable, and how much translation quality is affected by the errors of automatic acquisition.

This thesis discusses the following three tasks in order to clarify the features of corpus-based machine translation.

1. Proposing a knowledge acquisition method from bilingual corpora.

2. Applying the acquired knowledge to an actual MT engine and measuring the MT quality.

3. Identifying the inherent problems of the corpus-based MT that decrease MT quality and proposing solutions.

The first task is the most important for realizing a corpus-based MT, and it has been studied by many researchers. In most cases, they evaluate methods from the viewpoint of “whether the acquired knowledge is correct or not.” However, this is not sufficient. The author believes that *the inherent problems of corpus-based MT are not clarified until sentences are translated with the acquired knowledge.* Based on the above belief, we not only propose a knowledge acquisition method but also measure translation quality. Moreover, the inherent problems that decrease translation quality are identified, and some solutions are proposed in this thesis.

This research was carried out under the following conditions.

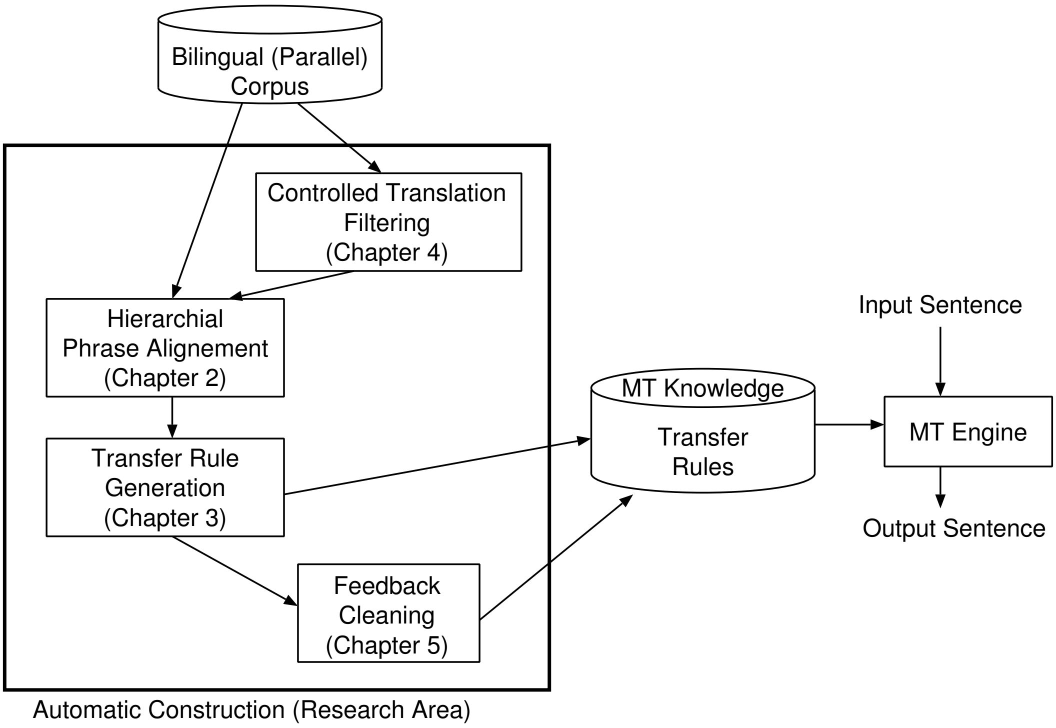

Bilingual (Parallel) Corpus
Controlled Translation Filtering (Chapter 4)
Hierarchial Phrase Alignement (Chapter 2)
Transfer Rule Generation (Chapter 3)
Feedback Cleaning (Chapter 5)
Automatic Construction (Research Area)
MT Knowledge
Transfer Rules
Input Sentence
MT Engine
Output Sentence

Figure 1.2. Research Area and Structure of Thesis

- The translation languages are from English to Japanese.

- The target domain is travel conversation. Thus, a bilingual corpus built for spoken language translation is used. Bilingual sentences between English and Japanese are aligned in advance.

- Example-based machine translation based on syntactic transfer (called transfer-based MT in this thesis) is used.

- The knowledge to be constructed includes the transfer rules, which represent the corresponding phrasal expressions between English and Japanese, for the transfer-based MT engine.

1.3 Structure of this Thesis

Figure 1.2 shows the structure of this thesis.

First, in Chapter 2, we propose a hierarchical phrase alignment (HPA) method between bilingual sentences, which is the nucleus of this work's automatic knowledge construction. This method automatically extract phrasal correspondences through parsing. It is difficult to parse sentences with high accuracy. This method resolves ambiguities by comparing the structures of a bilingual sentence. In addition, a partial tree sequence is output even if the parser cannot construct the tree structure of the entire sentence (Imamura, 2001).

Next, in Chapter 3, the HPA method proposed in Chapter 2 is applied to a large bilingual corpus, and transfer rules are constructed. In addition, the transfer rules are integrated into the MT engine, which is based on transfer driven machine translation (TDMT), and the translation quality is measured (Imamura, 2002).

In Chapter 4, before the automatic acquisition of MT knowledge, we attempt to resolve the problems of bilingual corpora themselves. Bilingual corpora do not contain only bilingual sentences that are appropriate for MT. They usually contain context/situation-dependent translations or multiple translations because one source sentence can be translated in various ways. In this chapter, we discuss bilingual sentences suitable for MT and then propose an automatic construction method that utilizes the features of the sentences (Imamura et al., 2003a).

In Chapter 5, we describe the post-processing done in the automatic acquisition stage. The acquired knowledge contains many redundant rules due to acquisition errors or translation variety in the corpora. If the redundant rules were removed, the translation quality could be improved. We propose a cleaning method for removing redundant rules in this chapter (Imamura et al., 2003b).

Finally, in Chapter 6, we conclude this thesis and discuss future directions of corpus-based machine translation.

Chapter 2

Hierarchical Phrase Alignment Harmonized with Parsing

2.1 Introduction

When building a machine translation system, we have to construct knowledge such as transfer rules manually. Therefore, automatic construction of machine translation knowledge is an effective way to reduce costs when we apply the system to other domains.

In this chapter, we propose a hierarchical phrase alignment method that aims to acquire translation knowledge automatically from bilingual sentences. Our method is especially suitable for spoken language translation, which contains many ungrammatical utterances. Here, phrase alignment (PA) refers to the extraction of equivalent partial word sequences between sentences of two languages. We use the term phrase alignment since these word sequences include not only words but also noun phrases, verb phrases, relative clauses, and so on. English and Japanese languages are used in this study.

For example, for the sentence pair:

E: *I have just arrived in New York.*
J: *Nyuyooku ni tsui ta bakari desu.*

the phrase alignment method should hierarchically extract the following word sequence pairs.

- *in New York* ↔ *Nyuyooku ni*
- *arrived in New York* ↔ *Nyuyooku ni tsui*
- *have just arrived in New York* ↔ *Nyuyooku ni tsui ta bakari desu*

We call these *equivalent phrases* in this thesis.

Equivalent phrases denote corresponding expressions between two languages. Therefore, they can be directly applied to example-based machine translation systems. In addition, because the phrases maintain hierarchical information, translation knowledge can be compressed by making hierarchical patterns in comparison with word sequences.

Some phrase alignment methods have already been proposed, such as those of Kaji et al. (1992), Matsumoto et al. (1993), Kitamura and Matsumoto (1995), Meyers et al. (1996), Watanabe et al. (2000), and Aramaki et al. (2001). The characteristics common to these previous methods are:

1. The methods employ parsers (phrase-structure analyzers or dependency analyzers) and word alignment (WA) results.
2. When they search for phrase correspondences, they only handle the final structures that the parser has output.
3. They handle only content word correspondences.

However, in the previous methods, the results of phrase alignment directly depend on the parsing accuracy. In particular, previous methods do not have any countermeasures against ungrammatical sentences failing the parsing. Therefore, these methods are not suitable for spoken language translation, which often involves handling ungrammatical sentences.

In this chapter, we propose a new method for phrase alignment that is harmonized with parsing. Our method resolves the ambiguity of parsing by comparing bilingual parse trees. When the parsing process fails, our method outputs partial phrase correspondences by combining partial parse trees. In addition, we increase the accuracy of phrase alignment by employing the word alignment results for not only content words but also functional words.

In the next section, we explain the basic method of hierarchical phrase alignment. Section 2.3 describes how to harmonize the basic method with parsing, Section 2.4 discusses suitable functions of word alignment for the phrase alignment, and Section 2.5 evaluates the performance, including comparisons with alternative methods.

2.2 Basic Method of Hierarchical Phrase Alignment

Generally speaking, most phrases in manual translation are translated into phrases of the same type, even if the language families are different. For ex-

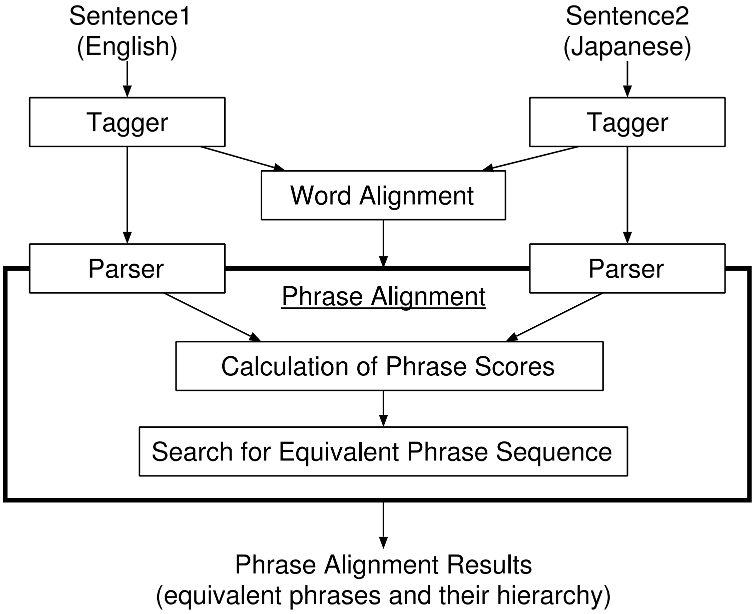


Figure 2.1. Flow of Hierarchical Phrase Alignment

ample, the English verb phrase "*arrive in New York*" is generally translated into the Japanese verb phrase "*Nyuyooku ni tsuku*".

Considering this feature, we assume that if partial word sequences of a bilingual sentence have the same semantic information, and if the phrase types are equal, the sequences can be regarded as equivalent phrases. We interpret this assumption as follows for computing:

Condition 1: "The same semantic information"
→ Words in the pair corresponded to no deficiency and no excess

Condition 2: "The same phrase types"
→ The phrases are of the same syntactic category

Therefore, this task is regarded as extracting phrases that satisfy the above two conditions. The following procedure shows the details of the extraction (Figure 2.1).

1. Tag and parse an English sentence and a Japanese sentence.

2. Extract corresponding words (called *word links*, represented as $WL(\text{English word}, \text{Japanese word})$) by word alignment. We assume

that W word links are extracted. Since many word alignment methods have been proposed elsewhere [1], here we do not discuss how they work.

3. Select i word links from among all links ($0 < i \leq W$), catch all of the syntactic nodes (non-terminal symbols), which include the links and exclude all other word links in the leaves, from the parsed English tree and Japanese tree.

4. Compare the syntactic categories of all English and Japanese nodes captured in Step 3. When identical node categories are found, regard the leaves of the nodes as equivalent phrases. If multiple candidates of a sentence or auxiliary verb phrase category are acquired, the candidate that covers the maximal area is selected. In other ambiguous cases, the candidate that covers the minimal area is selected.

5. Perform Steps 3. and 4. for all word link combinations.

Example (1): For example, suppose the English sentence "*I have just arrived in New York*" and its translation "*Nyuyooku ni tsui ta bakari desu*," which have two word links, i.e., $WL(\textit{New York}, \textit{Nyuyooku})$ and $WL(\textit{arrive}, \textit{tsui})$. When the parsing trees and word links are given as shown in Figure 2.2, the equivalent phrases are extracted as follows.

1. The syntactic node pair that contains only the word link $WL(\textit{New York}, \textit{Nyuyooku})$ (i.e., excludes the link $WL(\textit{arrived}, \textit{tsui})$) with nodes of the same syntactic category is retrieved. This finds the phrases `NP(1)` and `VMP(2)`.

2. Next, the syntactic node pair that contains only the word link $WL(\textit{arrived}, \textit{tsui})$ (i.e., excludes the link $WL(\textit{New York}, \textit{Nyuyooku})$) with nodes of the same syntactic category is retrieved. This finds the phrase `VP(3)`.

3. Finally, the node pairs that include both the word links $WL(\textit{New York}, \textit{Nyuyooku})$ and $WL(\textit{arrived}, \textit{tsui})$ with nodes of the same category are retrieved. This finds the phrases `VP(4)`, `AUXVP(5)`, and `S(6)`.

Therefore, the six equivalent phrases shown in Table 2.1 are extracted.

This is an example of two-word links. In the case of three-word links, the method retrieves phrases that include combinations of word links, such as those

[1] e.g., (Gale and Church, 1991; Melamed, 2000) and (Sumita, 2000)

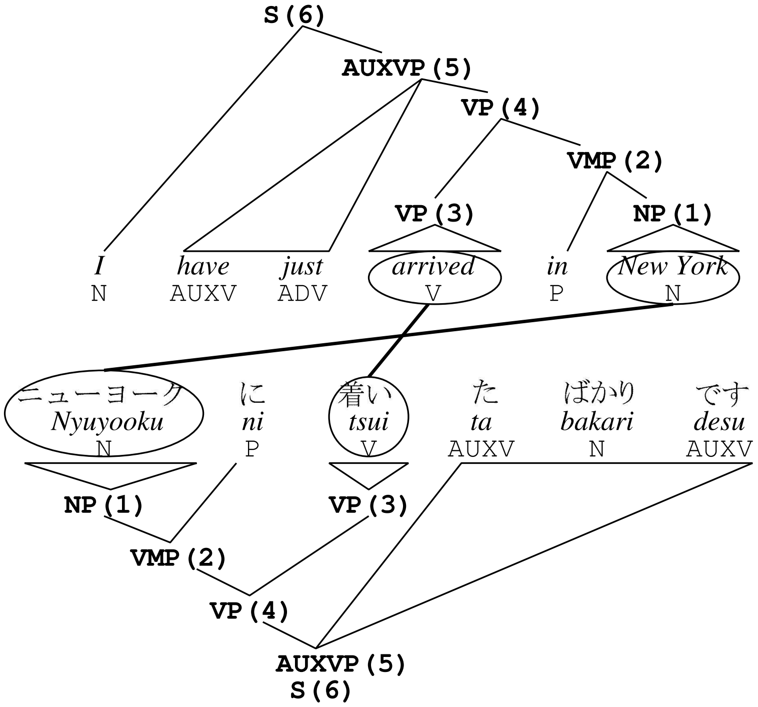


Figure 2.2. Example of Simple Translation
(upper and lower trees denote English and Japanese, respectively;
lines between languages denote word links)

including link 1, including links 1 and 2, and including all links. Equivalent phrases are extracted hierarchically.

Since the syntactic categories are different between English and Japanese, we classified the categories into seven types that are common to both languages, as shown in Table 2.2. Using this classification, we were able to compare different language categories.

Example (2): Even though a word link is available, the part-of-speech (POS) of a word is often different from that of its equivalent in a different language. If corresponding phrases are sought for using such word links without syntactic constraints, inappropriate translations will be extracted. However, the method described in this chapter only acquires phrases that have the same phrase type, so few unnatural short phrases are extracted as equivalents.

For example, consider extracting equivalent phrases from the English sentence "*Business class is fully booked*" and the Japanese sentence "*bijinesu-*

Table 2.1. Example of Phrase Alignment Results

Syntactic Category	English Phrase	Japanese Phrase
NP	*New York*	*Nyuyooku*
VMP	*in New York*	*Nyuyooku ni*
VP	*arrive*	*tsuku*
VP	*arrive in New York*	*Nyuyooku ni tsuku*
AUXVP	*have just arrived in New York*	*Nyuyooku ni tsui ta bakari desu*
S	*I have just arrived in New York*	*Nyuyooku ni tsui ta bakari desu*

Table 2.2. Type of Syntactic Categories

Phrase Type	Mark
Noun Phrase	NP
Verb Phrase	VP
VP with Auxiliary Verbs	AUXVP
Verb Modifier Phrase	VMP
Noun Modifier Phrase	NMP
Indepenent Phrase	INDP
Sentence	S
Other (language dependent phrase)	

kurasu wa yoyaku de ippai desu" (Figure 2.3). Even if the word links $WL(\textit{fully}/\mathrm{ADV}, \textit{ippai}/\mathrm{N})$ and $WL(\textit{booked}/\mathrm{V}, \textit{yoyaku}/\mathrm{N})$ are given, there is no pair of nodes that contains only one link between them and that are of the same syntactic category. However, there are VP(2) nodes that include both links and are of the same category. Therefore, the English phrase "*be fully booked*" and the Japanese phrase "*yoyaku de ippai desu*" are extracted as equivalents.

Example (3): An example of a non-literal translation is shown in Figure 2.4. In this example, because the English phrase "*fly*" is translated into the Japanese phrase "*hikoki de yuki* (go by plane)", they have no word links. However, the result of phrase alignment contains the English phrase "*fly to New York tomorrow*" and the Japanese phrase "*Nyuyooku ni asu hikoki de yuki*" as equivalents, so "*fly*" and "*hikoki de yuki*" are indirectly regarded as equivalents. Thus, this method is able to extract some non-literal equivalent phrases (i.e., non-word-by-word translation phrases) that lack word links.

Phrase alignment with a lack of word links is described in Section 2.4.2.

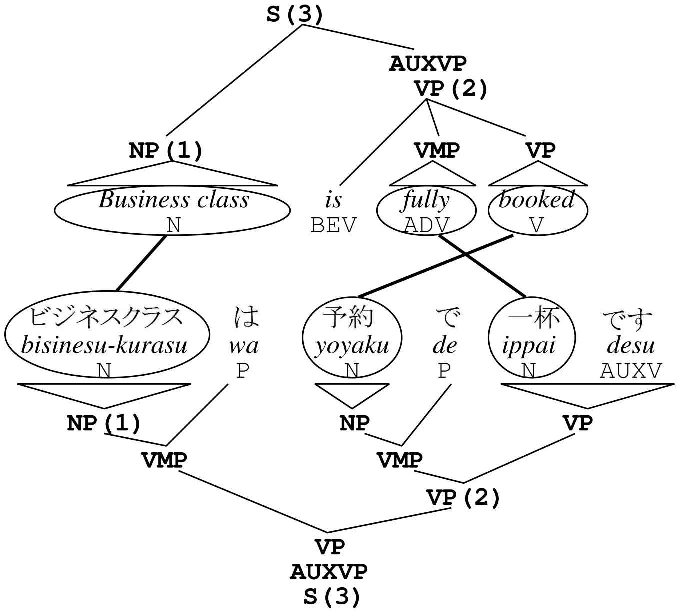


Figure 2.3. Example of Word Links for Words of Different POS

2.3 Phrase Alignment Harmonized with Parsing

The basic method described in Section 2.2 assumes that single parsing trees are given. However, phrase alignment results are directly affected by parsing results when they are processed after the determination of a single parsing tree.

For example, a bilingual sentence in which the parser cannot analyze its structure cannot be processed by phrase alignment. Moreover, incorrect parsing trees derive incorrect or insufficient phrase alignment results.

Parsing errors can be roughly classified into two types.

- Ambiguity:
  Parsing result contains multiple candidates, and the parser selects the wrong structure. In this case, the parsing result becomes incorrect.
- Failure in Constructing Parsing Tree:
  Because of the insufficient grammar (i.e., the lack of rewrite rules), the parser fails to create a tree that covers the whole sentence. In this case, the parser usually outputs no results.

Ambiguity always occurs when we parse a sentence. On the contrary, the

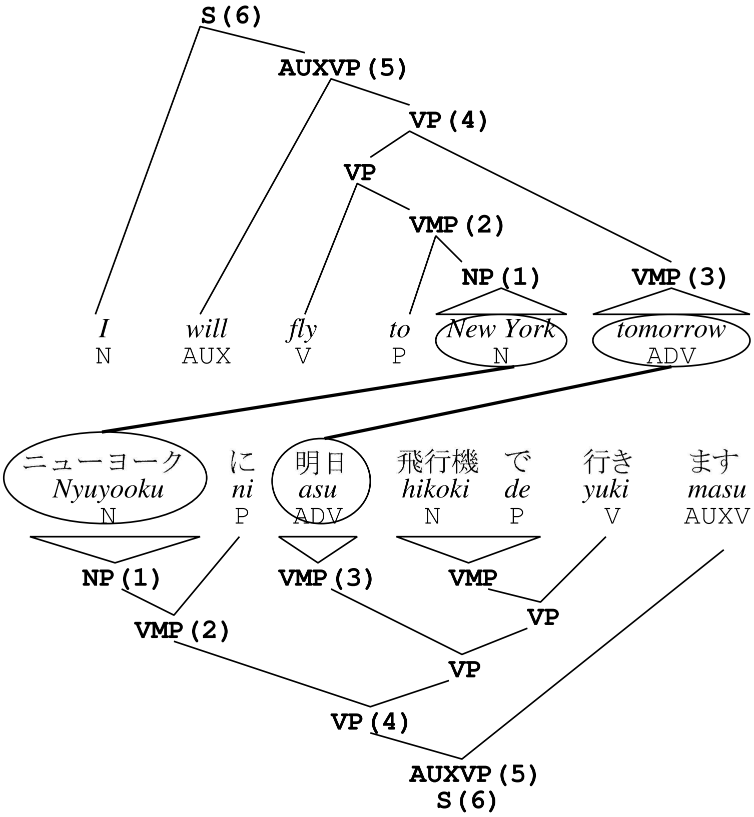


Figure 2.4. Example of Non-literal Translation

failure to construct a parsing tree can be suppressed if we prepare high-density grammar that include almost all rewrite rules. However, our target is spoken language translation. Even if we prepare high-density grammar, the parser cannot analyze sentences because spoken languages are often ungrammatical. In addition, machine translation has to analyze at least two languages. Corpora or language tools are different depending on the language. It is impossible to prepare parsers that never fail in any languages. Therefore, failure in constructing a parsing tree remains an unsolved problem.

Our proposed method solves these problems by harmonizing phrase alignment with the parsing by using the following two features and techniques.

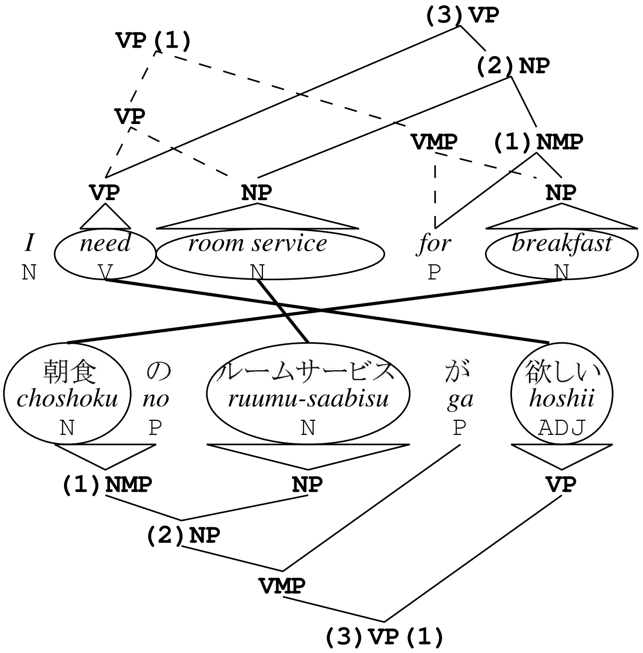


Figure 2.5. Example of Disambiguation for PP Attachment Modifyee

2.3.1 Disambiguation Using Structural Similarity between Languages

Some parsing ambiguities can be eliminated when the two languages are made to correspond. This type of disambiguation utilizes structural similarity (Kaji et al., 1992; Matsumoto et al., 1993).

For example, a prepositional phrase (PP) attachment modifyee in English is itself disambiguous when the equivalent Japanese phrase has only one structure. In Figure 2.5, the prepositional phrase "*for breakfast*" may modify either '*need*, ' and thus consist of `VP(1)` shown by the dotted line tree, or '*room service*, ' and consisting of `(2)NP` shown by the solid line tree. On the other hand, considering Japanese sentence structure, "*choshoku no*" definitely modifies '*ruumu-saabisu*' and consist of `(2)NP`. Therefore, "*room service for breakfast*" must be a noun phrase in the same way as Japanese.

This phenomenon shows that the conditions of ambiguity depend on the language, and some ambiguities can be resolved by intersecting their related conditions.

Thus, some disambiguation can be achieved by using an evaluation measure that outputs a high score when the structures of two languages become more similar.

We set the evaluation measure as follows:

- Form correspondences for all English and Japanese nodes with the two conditions described in Section 2.2.
- Select the structure that has the maximal number of corresponding nodes.

We call this measure *phrase correspondence score* in this thesis. For the solid line structure in Figure 2.5, `(1)NMP`, `(2)NP`, and `(3)VP` are evaluated as corresponding nodes. However, for the dotted line structure, only `VP(1)` is evaluated as a corresponding node in the same area. Therefore, the phrase correspondence score of the solid line structure is two greater than that of the dotted one, and the solid structure is selected.

Note that we assumed that there are no ambiguities in the word alignment result. If there are ambiguous word links (e.g., the same word appears twice in a sentence), the above evaluation measure can disambiguate them to some degree by searching for a word link combination that maximizes the phrase correspondence score.

2.3.2 Combination of Partial Trees

The phrase alignment method in this thesis utilizes a chart parser. Most parsers, including this one, output nothing when they fail to construct an entire parsing tree due to incomplete grammar (i.e., a lack of rewrite rules). However, they still keep partial trees in their agenda. Namely, correct partial tree candidates remain in the parser. If we combine these partial trees appropriately, we can recover the failure caused by incomplete grammar. This approach is especially effective for spoken language that contains many ungrammatical sentences (Takezawa and Morimoto, 1997). When we combine the partial trees, we have to check whether the partial trees are correct. The evaluation measure described in Section 2.3.1 is useful for examining this.

Naturally, when the parsing succeeds (i.e., a tree is derived from a whole sentence), the result should be preferred. Thus, we revise the measure in order to prefer the result constructed from the minimal number of partial trees. The evaluation measure is finally represented as follows.

1. Compare the nodes of a sentence pair, and extract the equivalent phrase candidates that maximize the phrase correspondence score.
2. Calculate the sum of the phrase scores in the partial tree sequence, and select the sequences that have the maximal score.

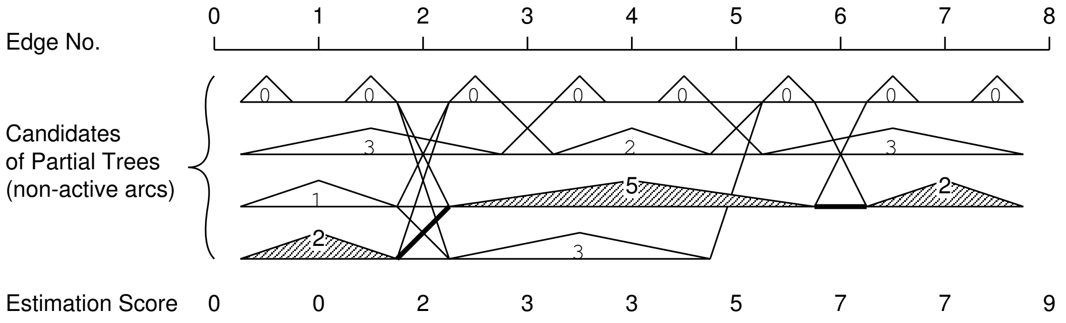


Figure 2.6. Example of Search for Partial Tree Combination (triangles denote partial trees, numbers in the triangles denote phrase correspondence scores, and mesh triangles denote the result that is searched for)

3. If multiple partial tree sequences are available, select the sequences that have the minimal number of partial trees.

However, an exponential quantity of time is required to examine all combinations of partial trees. In order to avoid this problem, we employ a *forward DP backward A^* algorithm* (Nagata, 1994), which is a two-pass search algorithm used for taggers.

The algorithm is described as follows. Figure 2.6 shows an example of search space for English partial trees. Note that the phrase correspondence score of each partial tree is calculated in advance. [2]

First, all partial trees of the English sentence are constructed into a lattice structure. When the forward search is applied, the maximum of the phrase correspondence score is computed from the start of the sentence to the edge $i(0 < i \leq$ the number of words $N)$ using dynamic programming. We call this the estimation score. Note that the paths are not recorded at this time. The estimation score ensures that at least one path exists from start to edge i.

In the backward search, the best combination of partial trees is searched for by using the A^* search algorithm. The estimation score is used as the heuristic function value of the A^* algorithm. Because the estimation score is the most accurate value of the heuristic function, the best path is searched for but no redundant paths are expanded while searching.

This algorithm searches for the best sequence of English partial trees and corresponding Japanese partial trees (i.e., the best sequence of equivalent phrases)

[2] In the current implementation, the phrase scores are calculated for all combinations of partial trees. Therefore, the calculation time is on the order of the number of English partial trees$*$ the number of Japanese partial trees.

without pruning. The searching time is nearly proportional to the word number.

Even though the English sequence is optimal, some Japanese partial trees overlap each other because a single English phrase corresponds to multiple Japanese phrases. Thus, when we apply the backward search, we have to check the areas of Japanese partial trees and do not expand the paths that include overlapping in the Japanese sequence. This means that some paths are discarded. However, the A^* search algorithm expands the next path that has the maximal score even if the current path is discarded. Therefore, the search result is guaranteed as optimal if the sum of the phrase correspondence score is maximal and the Japanese sequence does not have any overlapping.

2.4 Word Alignment for Phrase Alignment

2.4.1 Correspondence between Functional Words and Content Words

Functional words represent aspects, moods and so on, and they expand the variety of expressions. If we extract equivalent phrases while ignoring the functional words, phrases that are correct from the viewpoint of meaning but incorrect from the viewpoint of pragmatics would be extracted as the translation. Especially in Japanese, functional words are important because they represent tense.

Figure 2.7 shows an example. If there is no word link between '*after*' and '*iko*', `NP(1)` is extracted as an equivalent because there is only a word link $WL(\mathit{three}, \mathit{sanji})$ in the phrase. However, if there is the word link $WL(\mathit{after}, \mathit{iko})$, `NP(1)` is ignored because there is a deficient/excess link.

Word links between functional words or between a functional and content word make the constraint of Condition 1 tighter. Therefore, incorrect equivalent phrases can be ignored, and the accuracy of phrase alignment will consequently increase.

2.4.2 Relationship between Word Alignment Accuracy and Phrase Alignment

No word alignment (WA) method with 100% precision and recall rates has been proposed. Therefore, we should assume that word links include word alignment errors. With this in mind, which rate is more important for phrase alignment, precision or recall?

Under the condition of a 100% WA recall rate, a low WA precision rate would mean that the word links contain redundancy. As we described in Section 2.4.1, if

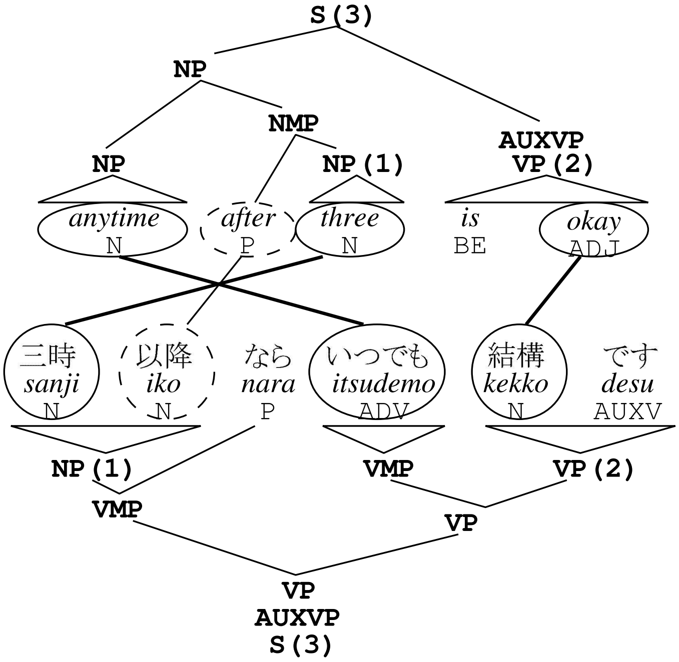


Figure 2.7. Example of Functional Word Link

the number of word links is increased, Condition 1 becomes tighter. Therefore, the number of equivalent phrases will decrease, but few incorrect equivalent phrases will be extracted.

On the other hand, in the case of a low WA recall rate (i.e., word links are insufficient), Condition 1 would become looser, and ambiguities, such as PP attachment modifyees, would increase. This, in turn, would cause incorrect equivalent phrases to be extracted. In addition, the number of equivalent phrases would decrease along with the reduction in the number of word link combinations.

Consequently, a word alignment method that has a high recall rate is more suitable for phrase alignment. In other words, word links should include most of the necessary links even though they include redundancy.

2.5 Trial Experiments

2.5.1 Experimental Settings

We conducted a number of experiments on phrase alignment, using 300 sentences containing basic travel expressions. These travel expressions were artificially created by humans imagining conversations. Therefore, they were not exact spoken language but did contain some ungrammatical sentences in comparison with written language. For example, some sentences were sequences of simple sentences or interjections without any conjunctions, such as "*You're very welcome, sir, please let me know if you have any problems, I'll be happy to help.*" Some sentences lacked particles in Japanese. The average number of words in a sentence was 8.95 in English and 8.81 in Japanese.

The details of the experimental settings are as follows:

- Tagging data was prepared by machine tagging with manual correction.
- Word links between content words were made manually. Links between functional words were made by referring to a translation dictionary.
- A basic bottom-up chart parser was used. The grammar was Context Free Grammar, which contained 286 English rules and 254 Japanese rules. The accuracies of the parser are shown in Table 2.3. [3] [4]

  The report of Charniak (2000) [5], who developed one of the English parsers, found that the labeled precision is 90.1% and the labeled recall is also 90.1% when the sentence length is less than 40 words. Compared with this work, our parsers have low labeled recall (i.e., many sentences could not parsed). This is because the grammars were made manually, so we could not cover all language phenomena by these grammars.
- In each experiment, 300 phrases were selected from the first candidate of the phrase alignment results, and bilingual evaluators evaluated them. One evaluator was a native English speaker, and another was a native Japanese speaker. The following three ranks were used for evaluation.

[3] We employed the evaluation metrics that were used in (Collins, 1997; Sekine and Grishman, 1995; Charniak, 2000).

$$\text{Labeled Precision} = \frac{\#\text{ of correct constituents in proposed parse}}{\#\text{ of constituents in proposed parse}}$$

$$\text{Labeled Recall} = \frac{\#\text{ of correct constituents in proposed parse}}{\#\text{ of constituents in treebank parse}}$$

$$\text{Crossing Brackets} = \#\text{ of constituents that violate constituent boundaries with a constituent in the treebank parse.}$$

[4] Grammars of English and Japanese were independently developed, but the numbers of outputs became 200 sentences for each grammar as a result.

[5] `ftp://ftp.cs.brown.edu/pub/nlparser/`

Table 2.3. Accuracies of Parsers Used in Experiments

	English	Japanese
# of Sentences	300	300
# of Output Sentences	200 (67%)	200 (67%)
# of Total Candidates (per sentence)	836 (4.18)	394 (1.97)
Labeled Precision	90.5%	93.1%
Labeled Recall	50.8%	52.6%
Crossing Brackets per Sentence	0.487	0.447
# of Sentences with Zero Crossing Brackets	144 (48%)	160 (53%)

A: Correct. It was a possible translation from the viewpoint of English to Japanese and Japanese to English.

B: Not wrong, but depends on context. It was a possible translation from the viewpoint of English to Japanese or Japanese to English.

C: Incorrect. It was a wrong translation in both English to Japanese and Japanese to English.

2.5.2 Effects of Harmonization

Differences in Accuracy among Phrase Alignment Methods First, we tested a variety of phrase alignment methods for the following three cases. The statistics of the results are shown in Table 2.4. The number of phrases in Table 2.4 denotes the number of equivalent phrases to which the evaluators assigned a rank.

Case 1: (Proposed Method) Extracting partial trees from the agenda while parsing, and searching for the best sequence with the maximal phrase correspondence score.

Case 2: Selecting the first parsing candidate only when whole sentence has been parsed, and processing the phrase alignment. This means that no countermeasures against ambiguity or failure of parsing were applied.

Case 3: Using all of the parsing candidates only when the entire sentence has been parsed, processing the phrase alignment for all combinations of candidates, and selecting the best one with the maximal phrase correspondence score. This means that ambiguities are resolved by using the phrase correspondence scores. Compared with Case 2, the effect of the phrase score

Table 2.4. Number of Equivalent Phrases and Accuracies

	Case	# of Output Sentences	# of Eq. Phrases (per output)	Accuracy of Eq. Phrase		
				Rank	# of Phrases	Ratio
Proposed Method	Case 1	296	1,676	A	248 + 269	86.2%
			(5.66)	B	30 + 5	5.8%
				C	22 + 26	8.0%
Alternative Phrase Alignment Methods	Case 2	176	726	A	249 + 270	86.5%
			(4.13)	B	30 + 8	6.3%
				C	21 + 21	7.0%
	Case3	177	822	A	264 + 267	88.5%
			(4.64)	B	18 + 3	3.5%
				C	18 + 30	8.0%
Variable Word Links	Case 4	295	1,703	A	240 + 258	83.0%
			(5.77)	B	31 + 4	5.8%
				C	29 + 36	10.8%
	Case 5	276	1,018	A	245 + 266	85.2%
	WA Prec.: 50%		(3.69)	B	17 + 0	2.8%
	WA Recall: 100%			C	38 + 31	11.5%
	Case 6	272	1,147	A	209 + 230	73.2%
	WA Prec.: 100%		(4.22)	B	21 + 4	4.2%
	WA Recall: 50%			C	70 + 66	22.7%

becomes clearer. Moreover, compared with Case 1, the effect of the partial tree combination becomes clearer.

First, in Case 1, the accuracy of the equivalent phrases (we only consider Rank A) was about 86.2%.

Comparing Cases 2 and 3, the number of equivalent phrases increased in Case 3. The phrase correspondence score selects a candidate that has many correspondences in the tree, so the number increased. However, the accuracy of equivalent phrases was almost the same. The reason for the same accuracy is that this method essentially makes few incorrect correspondences because the equivalent phrases have to match their syntactic categories. Thus, incorrect parsing trees were ignored when extracting equivalent phrases.

Comparing Cases 1 and 3, the number of equivalent phrases nearly doubled because almost all sentences were analyzed in Case 1. However, accuracy was almost the same. Therefore, the phrase scores work appropriately during a combination of partial trees. Such a combination is especially effective for spoken languages because we can obtain equivalent phrases from ungrammatical sentences.

2.5.3 Influence of Word Alignment

In order to examine the influence of word alignment, we conducted tests using various word links. All of the experiments used the Case 1 phrase alignment method. The results are shown in Table 2.4.

Case 4: Word links are limited to content words. Compared with Case 1, the effect of the functional word links becomes clearer.

Case 5: A variable WA precision rate under a fixed WA recall rate. The word links in Case 1 were regarded as perfect, and the precision rate was changed from 50% to 100%. The purpose of this experiment was to measure the influence of redundant word links. The redundant word links were made by selecting word pairs randomly that are not included in the original links.

WA precision and recall rates are represented by the following equations:

$$\text{WA precision rate} = \frac{N_{org}}{N_{org} + N_{red}} \tag{2.1}$$

$$\text{WA recall rate} = \frac{N_{org} - N_{eli}}{N_{org}} \tag{2.2}$$

Where N_{org} is the number of original word links, N_{red} is the number of redundant links and N_{eli} is the number of eliminated word links.

Case 6: A variable WA recall rate under a fixed WA precision rate. The recall rate was changed from 50% to 100%. The purpose of this experiment was to measure the influence of insufficient word links. The insufficient links were made by eliminating links randomly.

Effect of Function Word Links: Comparing Cases 1 and 4, the number of extracted equivalent phrases slightly increased in Case 4, and the accuracy of the equivalent phrases slightly decreased. In order to verify this result, we re-evaluated 50 phrases that appeared only in Case 1 and 50 phrases that appeared only in Case 4 by one Japanese native speaker. The numbers of Rank A were 36 (72%) in Case 1 and 14 (28%) in Case 4. Therefore, it was verified that the accuracy of equivalent phrases increase when the result of word alignment includes functional links.

Influence of Word Alignment Accuracy: Figure 2.8 shows the number of extracted equivalent phrases in Cases 5 and 6 according to variable WA precision and recall rates. In both cases, the number of phrases decreased similarly

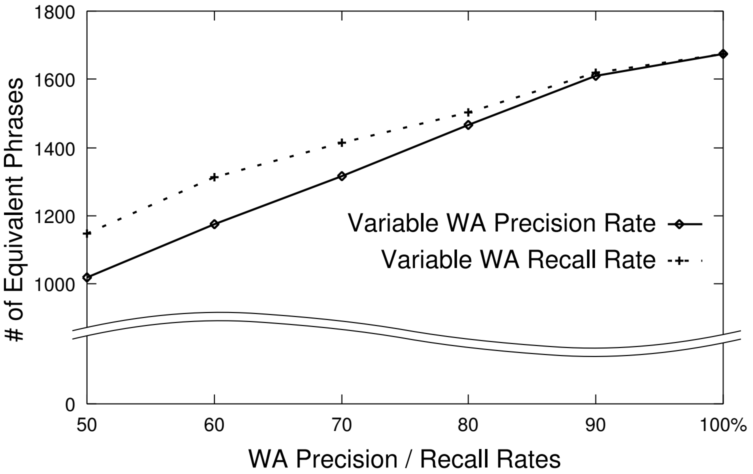


Figure 2.8. Number of Extracted Equivalent Phrases According to Word Alignment Accuracies

according to the decrease in WA accuracy, but this was slightly affected by the WA precision rate.

On the other hand, Table 2.4 indicates that when the WA precision rate decreased, the accuracies of the equivalent phrases were nearly equal, but when the WA recall rate decreased, accuracy clearly decreased. Therefore, as we described in Section 2.4.2, the WA recall rate affects the phrase alignment accuracy with greater sensitivity. In other words, the easiest way to increase phrase alignment accuracy is to provide as many word links as possible.

2.5.4 Examples of Equivalent Phrases

Table 2.5 shows examples of equivalent phrases extracted by the proposed method.

(A) is an example where English and Japanese parsing both failed because the sentence is a sequence of interjections and simple sentences without conjunctions. Even if the parsing fails, our method can output equivalent phrases from simple sentence pairs, including their low-level structures. Note that the incorrect equivalent phrases, numbers 7 and 8, were extracted due to an incorrect parsing result (the correct English structure must be [*your* [*passport and ticket*]], but our parser made the incorrect structure [[*your passport*] *and* [*ticket*]]). The structures of parallel phrases are ambiguous in both English and Japanese, so not all ambiguities can be resolved by comparing two languages.

(B) is an example that lacked a Japanese particle. In this case, only Japanese

Table 2.5. Examples of Equivalent Phrases Extracted by Proposed Method

(A) Sequence of Interjections and Simple Sentences

English: *All right, I understand, here is your passport and ticket.*
Japanese: *ookei, wakari mashi ta, hai, anata no pasupooto to koukuuken desu.*

No.	Syn. Cat.	English Phrase Japanese Phrase
1	S	*I understand* わかりました *wakari mashi ta*
2	AUXVP	*understand* わかりました *wakari mashi ta*
3	VP	*understand* わかり *wakari*
4	S	*here is your passport and ticket* あなたのパスポーと航空券です *anata no pasupooto to koukuuken desu*
5	AUXVP	*is your passport and ticket* あなたのパスポートと航空券です *anata no pasupooto to koukuuken desu*
6	VP	*be your passport and ticket* あなたのパスポートと航空券です *anata no pasupooto to koukuuken desu*
7	NP	*your passport* あなたのパスポート *anata no pasupooto*
8	NP	*ticket* 航空券 *koukuuken*

(B) Sentence that Lacks Case Particles

English: *Please retrieve my coat.*
Japanese: *azuke ta kooto, dashi te kudasai*

No.	Syn. Cat.	English Phrase Japanese Phrase
1	S	*please retrieve* 出してください *dashi te kudasai*
2	VP	*retrieve* 出し *dashi*
3	NP	*my coat* コート *kooto*

parsing failed, but our method could extract three partial equivalent phrases. The reason why phrase number 1 was extracted as S is that the Japanese parser could not make a larger tree. However, if we assume that the phrase "*azuke ta kooto*" was used as an independent clause, "*please retrieve*" should correspond to "*dashi te kudasai.*"

2.6 Related Work

The following research papers, which hierarchically acquired phrasal correspondences by matching between the parsing trees of two languages, have been proposed.

First, a research work based on the phrase structure, Kaji et al. (1992), proposed a method that extracts the node correspondence between phrase structures of two languages from the result of word alignment. Our research is based on this method. However, Kaji et al. (1992) did not consider the syntactic category constraint. Therefore, if the POS's of words, which are the edges of a word link, are different, unnatural short phrases would be extracted as equivalent.

Other research works have been based on the dependency structure (Matsumoto et al., 1993; Kitamura and Matsumoto, 1995; Meyers et al., 1996; Yamamoto and Matsumoto, 2000; Watanabe et al., 2000; Aramaki et al., 2001). Nodes in the dependency structure represent minimal units of syntactic phrases. Therefore, some phrasal correspondence can be extracted without syntactic category information. However, we believe this approach has the same problem as that of Kaji et al. (1992).

Wu (1995) proposed a synchronized algorithm that simultaneously extracts phrasal correspondences while parsing two languages. This algorithm requires synchronized grammar rules for parsing, in which single words (terminal symbols) of two languages are corresponded in advance. In other words, only word alignment is necessary. This is suitable for literal translations, where almost all words in the two languages correspond to each other. However, we assume that this is not suitable for spoken languages that contain non-literal translations because the syntactic constraint is weak.

No method have given countermeasures against failed parsing. The accuracy of parsing decreases when the input sentence is out of the domain for which the grammar has been designed. Our method can extract equivalent phrases by combining partial parsing results, even though it utilizes parsers with low coverage of the grammar. Therefore, our method has the advantage of being able to extract many equivalent phrases when it is applied to other domains.

From the viewpoint of the usage of word alignment, Kaji et al. (1992) and

Watanabe et al. (2000) explicitly applied the result of word alignment. On the other hand, Matsumoto et al. (1993), Kitamura and Matsumoto (1995), and Meyers et al. (1996) introduced similarity scores of words and utilized them to calculate the structural similarity.

The research of Yamamoto and Matsumoto (2000) has the feature of not requiring word alignment. This method creates candidates of equivalent phrases and decides phrasal correspondences by the best-first algorithm using a weighted dice coefficient. Our method does not utilize any statistical information, but assumes that introducing it would further increase the accuracy of equivalent phrases.

2.7 Conclusions

In this chapter, we proposed a hierarchical phrase alignment method that is harmonized with parsing. The method uses a phrase correspondence score to evaluate syntactic structural similarity. With this method, we could carry out disambiguation and partial tree combination.

In particular, the proposed method could extract about twice as many equivalent phrases as independent parsing, even if we utilized parsers with low labeled recall. The accuracy of the extracted equivalent phrases was about 86%, and almost no deterioration was observed. In addition, we showed that the accuracy of phrase alignment increases when the result of word alignment contains functional word links.

Since the proposed method has greater sensitivity to a lack of necessary word links, it extracts better equivalent phrases when it uses a word alignment method with a high recall rate.

In the next chapter, we will construct translation knowledge from the extracted equivalent phrases and apply the knowledge to a transfer-based machine translation system.

Chapter 3

Application of Translation Knowledge Acquired by Hierarchical Phrase Alignment for Transfer-based MT

3.1 Introduction

Translation knowledge is necessary for machine translation (MT) systems. Automatic construction of translation knowledge is an effective way to reduce costs when applying a system to other task domains.

Statistical machine translation methods (e.g., (Brown et al., 1993)) automatically acquire statistical models, which are considered elements of translation knowledge, so little cost is necessary. However, in most cases, these methods are applied to the same language families, such as English and French. In the case of different families, the translation quality is still unclear.

A hierarchical phrase alignment method (HPA) has been proposed in Chapter 2. This method hierarchically extracts equivalent phrases from a sentence-aligned bilingual corpus even though they belong to different language families. Kaji et al. (1992), Yamamoto and Matsumoto (2000), and Meyers et al. (2000) have also proposed methods to acquire translation knowledge automatically. They have evaluated the knowledge, but there are few examples in which the translation quality was evaluated when the entire knowledge was applied to translation systems (Menezes and Richardson, 2001). This comprehensive level of quality should be measured on an actual translation system to judge whether the acquired knowledge is useful from a practical point of view.

In this chapter, transfer rules, which are a kind of translation knowledge, are acquired automatically by hierarchical phrase alignment and integrated into a transfer-based MT system, and then the resulting translation quality is evaluated. Through the integration, the problem of ungeneralized rules contained within the knowledge became clear. Because this problem cause bad translations or increase ambiguities, it become obvious that the knowledge needed to be cleaned. The languages studied here is from English to Japanese.

3.2 Overview of Hierarchical Phrase Alignment

Details were described in Chapter 2.

3.2.1 Basic Method

Phrase alignment refers to the extraction of equivalent partial word sequences between bilingual sentences. We use the term phrase alignment since these word sequences include not only words but also noun phrases, verb phrases, relative clauses, and so on.

For example, when the following bilingual sentence is given,

English: *I have just arrived in New York.*
Japanese: *Nyuyooku ni tsui ta bakari desu.*

the phrase alignment should extract the following word sequence pairs.

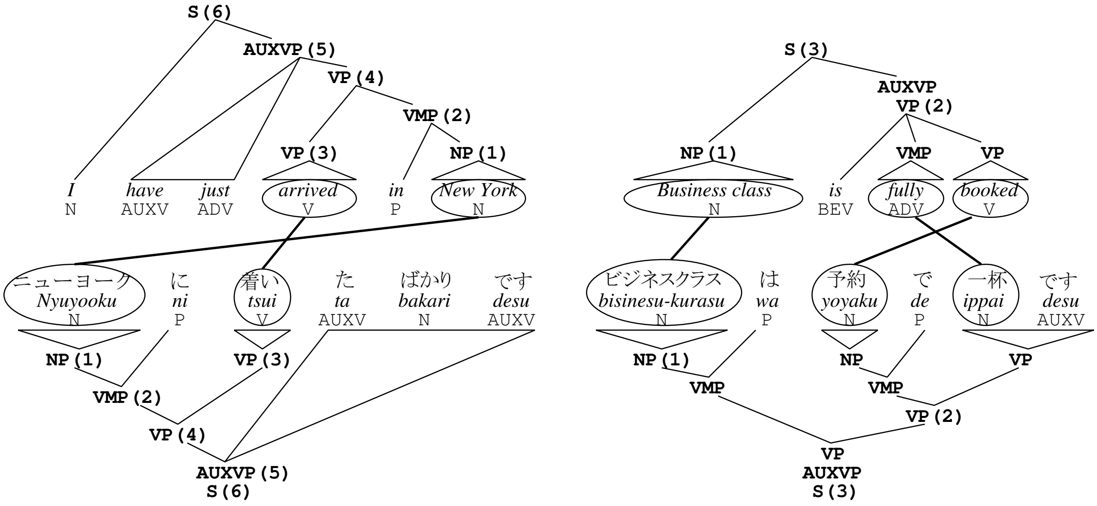
- *in New York* ↔ *Nyuyooku ni*
- *arrived in New York* ↔ *Nyuyooku ni tsui*
- *have just arrived in New York* ↔ *Nyuyooku ni tsui ta bakari desu*

We call these *equivalent phrases* in this thesis and defined this task as extracting phrases that satisfy the following two conditions.

Condition 1 (Semantic constraint):
Words in the phrase pair correspond to no deficiency and no excess.

Condition 2 (Syntactic constraint):
The phrases are of the same syntactic category.

In order to extract phrases that satisfy two conditions, corresponding words (called *word links*, represented as $WL(word_e, word_j)$) are first extracted by word alignment. Next, the sentence pair is parsed respectively, and phrases and their syntactic categories are acquired. Finally, the phrases, which include some word



(a) Example of Simple Translation (b) Example of Different POS Translation

Figure 3.1. Examples of Hierarchical Phrase Alignment
(Upper and lower trees denote English and Japanese, respectively;
lines between languages denote word links.)

links, exclude other links, and are of the same syntactic categories, are regarded as equivalent.

For example, in the case of Figure 3.1(a), NP(1) and VMP(2) are regarded as equivalent because they only include $WL(\textit{New York}, \textit{Nyuyooku})$, and are of the same syntactic category. In the case of $WL(\textit{arrived}, \textit{tsui})$, VP(3) is regarded as equivalent, and in the case of both word links, VP(4), AUXVP(5), and S(6) are regarded as equivalent. Consequently, six equivalent phrases are extracted hierarchically.

Even though word links are available, the part-of-speech (POS) of the words is sometimes different in different languages, as shown in the second example in Figure 3.1(b). In this case, the phrases that contain only $WL(\textit{fully}, \textit{ippai})$ or only $WL(\textit{booked}, \textit{yoyaku})$ are not regarded as equivalent because of the syntactic constraint, and VP(2) nodes are extracted first. Thus, few unnatural short phrases are extracted as equivalent.

3.2.2 Increasing Robustness

The problem in the above method is that the result of the phrase alignment directly depends on the parsing result. We solved this problem by using the following features and techniques, and partial correspondences were extracted even though parsing failed. In the experiment in Chapter 2, about twice as many equivalent phrases were extracted compared with the basic method and almost no deterioration was observed.

Disambiguation Using Structural Similarity: As Kaji et al. (1992) and Matsumoto et al. (1993) showed, some parsing ambiguities can be resolved when the two languages are made to correspond. This disambiguation utilizes structural similarity. For example, a PP attachment in English is ambiguous as to whether it modifies a noun or a verb, but this is nearly always definite in Japanese. Hence, when the attachment is assumed to modify the same word, the ambiguity is resolved. Accordingly, the structures between the two languages become similar.

We employ a '*phrase correspondence score*' to measure structural similarity. This measure is calculated by counting the phrases that satisfy the above two conditions, and the parsing candidate that has the maximal score is selected.

Combination of Partial Trees: Partial parsing is an effective way to avoid a lack of grammar or to parse ungrammatical sentences. It is used to combine partial candidates in the parser. Therefore, a criterion as to whether the part is valid or not is necessary for the combining process. We utilize the phrase correspondence score as the criterion, and a partial tree sequence that maximizes the sum of the phrase correspondence scores is searched for. The forward DP backward A^* search algorithm (Nagata, 1994) is employed to speed up the combination.

3.2.3 Advantage of HPA

The phrase alignment result by this method maintains correspondent parsing trees and hierarchical information, so it is especially suitable for transfer-based MT systems (i.e., MT systems using syntactic transfer methods).

Moreover, a characteristic of this method is the introduction of a syntactic constraint (Condition 2). [1] There are two effects of the syntactic constraint. One is that few unnatural short phrases are extracted, as described above. The other is that it is easy to construct transfer rules because the phrases can be grammatically replaced.

In other words, suppose that an equivalent phrase replaces another one that is extracted from another sentence. If the source phrase and the target phrase are in the same syntactic category, the resulting synthesized sentence is appropriate. On the other hand, if they are in different categories, the source or target sentence

[1] The methods of Yamamoto and Matsumoto (2000) and Meyers et al. (2000) do not use syntactic categories. Alternatively, dependency structures are utilized. Chunks and relationships may be substituted for categories. However, this approach is not declarative.

becomes grammatically inappropriate. The syntactic constraint suppresses such inappropriate substitution. This is a particular advantage for translation between different language families, since this phenomenon appears more frequently in such case than in translation between languages of the same language family.

3.3 Transfer Driven Machine Translation (TDMT)

The Transfer Driven Machine Translation system, or TDMT (Furuse and Iida, 1994; Sumita et al., 1999), used here is an example-based MT system (Nagao, 1984) based on the syntactic transfer method (called transfer-based MT). The following sections describe the overview of TDMT focusing on the transfer module.

3.3.1 Transfer Rules

Transfer rules represent the correspondence between source language expressions and target language expressions. They are the most important kinds of knowledge in TDMT. Examples are shown in Figure 3.2 that include the preposition '*at.*' In this rule, source language information is constructed by a source pattern and its syntactic category. The source pattern is a sequence of variables and constituent boundaries (functional words or part-of-speech bigram markers). The each variable is restricted by a syntactic category using daughter rules. Namely, source language information is equivalent to Context Free Grammar such that the right side of each rewrite rule absolutely contains at least one terminal symbol.

Target patterns are similarly constructed with variables and constituent boundaries, but they do not have POS bigram markers. In addition, each rule has source examples, which are instances of variables. The source examples are headwords acquired from training sentences. For instance, the first rule of Figure 3.2 means that the English phrase "*present at (the) conference*" was translated into the Japanese phrase "*kangi* (conference) *de happyo-suru* (present)."

3.3.2 Translation Process

At the time of translation, the source sentence is parsed using source patterns. Then, the target structure, which is mapped by target patterns, is generated (Figure 3.3). However, as shown in Figure 3.2, one transfer rule has multiple target patterns. In order to select an appropriate target pattern, semantic distances (node distances on the thesaurus; refer to (Sumita and Iida, 1991)) are calculated

Syn. Cat.	Source Pattern		Target Pattern	Source Example
`VP`	$\mathrm{X}_{\mathtt{VP}}$ *at* $\mathrm{Y}_{\mathtt{NP}}$	⇒	Y' *de* X' Y' *ni* X' Y' *wo* X'	((*present*, *conference*) ...) ((*stay*, *hotel*), (*arrive*, *p.m*) ...) ((*look*, *it*) ...)
`NP`	$\mathrm{X}_{\mathtt{NP}}$ *at* $\mathrm{Y}_{\mathtt{NP}}$	⇒	Y' *no* X'	((*man*, *front desk*) ...)

Figure 3.2. Examples of Transfer Rules in which the Constituent Boundary is '*at*'

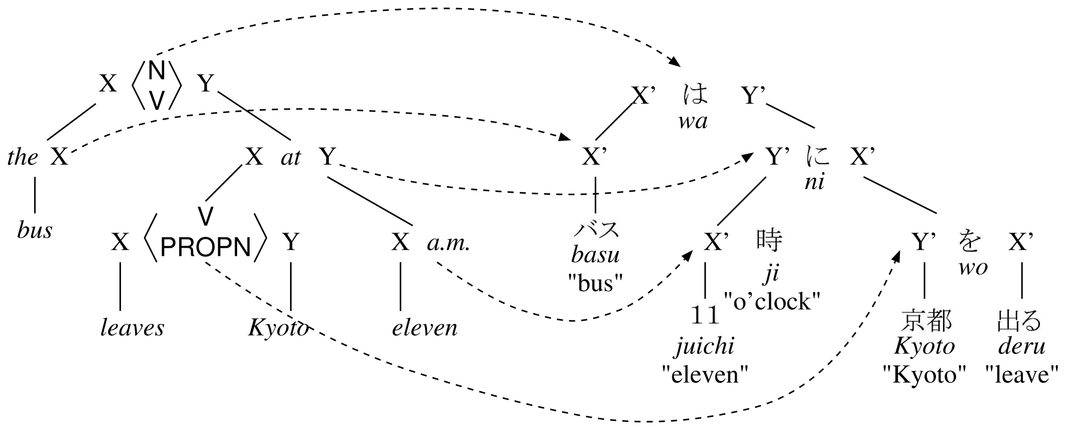


Figure 3.3. Example of TDMT Transfer Process

between the source examples and the daughter headwords of the input sentence, and the target pattern that has the nearest example is selected. Therefore, each rule also has head information.

For example, when the input sentence "*The bus leaves Kyoto at eleven a.m.*" is given, the source pattern (X *at* Y) is used. Then, the headword of the variable X is '*leave*,' and Y is '*a.m.*' According to the semantic distance calculation, the source example (*arrive*, *p.m.*) is the nearest. Therefore, the target pattern (Y' *ni* X') is selected. The semantic distance is also applied to parsing disambiguation.

3.3.3 Content Word Selection

Functional words are translated by the above process. In the case of content words, TDMT generates a default translation at leaf variables by referring to a translation dictionary. However, a single word is often translated into different words in different contexts. For example, in the case of the English phrase "*leave Kyoto*," '*leave*' should be translated into '*deru* (go out).' On the other hand, in the case of "*leave my wallet on the table*," '*leave*' should be translated into '*okisaru* (put and go).'

Content word selection is achieved in two ways. One is by using local dictionaries, which are translation dictionaries created for each target pattern. When an instantiated variable of a source pattern equals an example, the system refers to the local dictionary and generates the translated word (Yamada et al., 1998). Another way is by embedding content words that can generate different translations into the source and target pattern in advance.

3.4 Application of HPA Results for TDMT

In this section, we describe how to generate TDMT transfer rules from the results of HPA and the problems of this method.

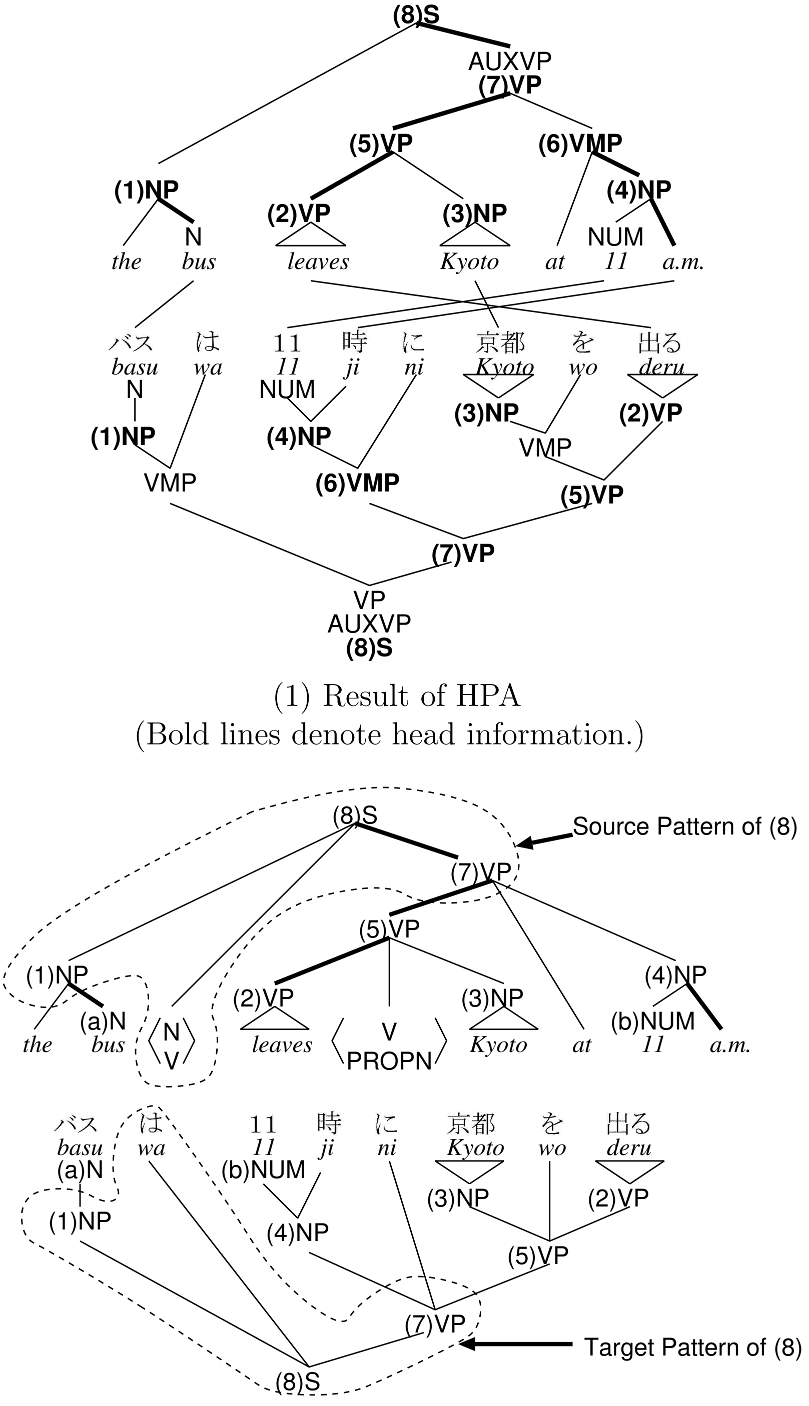
3.4.1 Transfer Rule Generation

The transfer rules described in Section 3.3 are constructed by source patterns that include their syntactic category, target patterns, source examples, head information, and local dictionaries. They are generated as follows from the HPA results (Figure 3.4).

1. First, the result of HPA is transformed into a structure that can construct transfer rules.

    - If an input word sequence includes continuous content words, insert a bigram marker in the intermediate of the content words. The bigram marker is an artificial word, which works as a functional word when the translator parses the input sentence.
    - If the edges of a word link are content words and are of the same POS types, a new word-level correspondence is added. A function of this correspondence is to translate unseen words by referring to the translation dictionary.

      In Figure 3.4, the correspondences `(a)N` and `(b)NUM` are supplied from the word links $WL($*bus*/N, *basu*/N$)$ and $WL($*11*/NUM, *11*/NUM$)$.
    - If variables are continuous when the source pattern is generated, the correspondence is removed because TDMT does not accept the series of variables. In Figure 3.4, `(6)VMP` is removed.
    - All nodes that do not have correspondence are removed except for the top node.



(1) Result of HPA
(Bold lines denote head information.)

(2) Tree Structure after Transformation

Figure 3.4. Example of Transfer Rule Generation

Syn. Cat.	Source Pattern		Target Pattern	Source Example
(8)S	$X_{\mathtt{NP}}$ <N-V> $Y_{\mathtt{VP}}$	⇒	X' *wa* Y'	(*bus*, *leave*)
(7)VP	$X_{\mathtt{VP}}$ *at* $Y_{\mathtt{NP}}$	⇒	Y' *ni* X'	(*leave*, *a.m.*)
(5)VP	$X_{\mathtt{VP}}$ <V-PROPN> $Y_{\mathtt{NP}}$	⇒	Y' *wo* X'	(*leave*, *Kyoto*)
(1)NP	*the* $X_{\mathtt{N}}$	⇒	X'	(*bus*)
(4)NP	$X_{\mathtt{NUM}}$ *a.m.*	⇒	X' *ji*	(*11*)

Figure 3.5. Example of Generated Rules from the Sentence "*The bus leaves Kyoto at 11 a.m.*"

2. Next, source patterns, target patterns, source examples, head information, and local dictionaries are created as follows.

    - Source patterns and target patterns are generated from the correspondences. The patterns are generalized by regarding daughter corresponding nodes as variables.
    - Head information is acquired from grammar, and source examples are identified by tracing the parsing tree to the head branch.
    - Local dictionaries are created by word links and by extracting leaf equivalent phrases in which the source phrase contains only a word.

In addition, because the inputs of phrase alignment are aligned sentences, sentence correspondences are added to the phrase alignment results as equivalent when the top nodes of the trees don't have the correspondence.

When the result of HPA is given as shown in Figure 3.4, five rules are generated as shown in Figure 3.5. Note that rules are not generated from the correspondences (2)VP, (3)NP, (a)N, and (b)NUM, because they are output to the local dictionaries.

3.4.2 Problems of Generated Transfer Rules

Even after transfer rules are generated, they may contain many incorrect or redundant (ungeneralized) rules. The reasons for this are classified as follows (Table 3.1).

(1) Reasons for Incorrect Translation

(1-a) **Context/Situation-Dependent Translation**
Omissions or additional words are contained in rules due to context or

Table 3.1. Causes of Incorrect/Redundant Rules

	Incorrect Translation	Ambiguity
Problems in Corpora	(1-a) Context/Situation Dependent Translation	(2-a) Multiple Expressions
Problems of HPA	(1-b) Incorrect Phrase Alignment	(2-b) Lack of Correspondence

situation-dependent equivalent phrases. For instance, the determiner '*the*' is not generally translated when English is translated into Japanese. However, when a human translator cannot semantically identify the following noun, a determinant modifier such as '*watashi-no* (my)' or '*sono* (its)' is supplied. These rules depend on the context, so if they are used in the different context, the translation will be incorrect.

(1-b) **Incorrect phrase alignment**
This causes the incorrect transfer rules not only in themselves but also in parent rules that have variables instantiated by the result.

For example, in Figure 3.1(b), suppose that an incorrect pair of the English phrase "*book*" and the Japanese phrase "*yoyaku de ippai desu*" are extracted as equivalent. This result will deliver the incorrect transfer rules (*book*) ⇒ (*yoyaku de ippai desu*) and (X *is fully* Y) ⇒ (X' *wa* Y').

The experiment described in Chapter 2 shows that 6% of the phrases were context dependent and 8% were incorrect even if word alignment was carried out by hand.

(2) Reasons for Correct Translation but Redundant Rules

(2-a) **Multiple Expressions**
The corpora usually contain a variety of translations even for a single source sentence because a sentence can be translated many ways. For example, in the corpus used for the experiments of Section 3.5, the English sentence "*How can I get there?*" is translated into thirty Japanese sentences. These translations cause various rules. However, they can be unified, so most rules will be unnecessary.

(2-b) **Lack of Correspondence**
When the result of HPA partially lacks correspondence, rules in which the variables are instantiated in advance are generated.

Table 3.2. Statistics of Corpus

	English	Japanese
# of Sentences	125,579	
# of Total Words	721,848	774,711
# of Different Words	9,945	14,494
# of Equivalent Phrases	404,664	
(including Sentence Correspondence)	(463,869)	
# of Different Patterns	56,851	53,317

For example, if the correspondence `VP(2)` is missing in Figure 3.1(b), the transfer rule (X *is fully booked*) ⇒ (X *wa yoyaku de ippai desu*) will be generated from `S(3)`. These rules are correct but clearly ungeneralized.

Meyers et al. (2000) referred to this problem as an explosive number of rules and decreasing translation speed. They tried to solve it by selecting rules based on the frequency during translation. TDMT performs rule selection based on the semantic distance, so the translation speed decreases only slightly even if there are many rules. On the other hand, because TDMT does not employ frequency, low-frequency rules of Type (1) cause incorrect translation. Therefore, they should be cleaned in advance.

The experiment in the next section compares the translation quality among different cleaning methods.

3.5 Evaluation

English to Japanese translation is evaluated in this chapter.

3.5.1 Experimental Settings

Corpus for Rule Generation We built a collection of Japanese sentences and their English translations based on expressions that are usually found in phrasebooks for foreign tourists (Takezawa et al., 2002; Kikui et al., 2003). We used about 125K sentences in the corpus for this experiment, and the basic statistics are shown in Table 3.2. The numbers of different patterns in Table 3.2 denotes the numbers of different source (English) and target (Japanese) patterns, respectively.

Evaluation Measure Each experiment used the same test set, which was composed of 508 sentences randomly selected in advance from the corpus and excluded from the training set. The evaluation was carried out by one Japanese native speaker. He/She evaluated the EJ translation into the following four ranks (Sumita et al., 1999) from the viewpoint of a user. In this chapter, we call (A+B+C) the *translation rate*.

(A) Perfect: no problem in either information or grammar.
(B) Fair: easy-to-understand with some unimportant information missing or flawed grammar.
(C) Acceptable: broken but understandable with effort.
(D) Nonsense: important information has been translated incorrectly.

Cleaning Methods We employed the following rule cleaning methods.

- **Baseline:**
  All rules were integrated into TDMT.

- **Cutoff by Frequency:**
  The frequency was counted for each source and target pattern pair, and transfer rules were generated only from high-frequency pairs in the same manner as in (Menezes and Richardson, 2001) experiment. In this experiment, the pairs that appeared more than two times were used.

- **χ^2 Test:**
  Considering that source and target patterns occur independently, the χ^2 test was performed. In this process, only high-frequency rules were tested in order to rely on the χ^2 value. That is, the co-occurrence frequency was over 40, or the co-occurrence frequency was over 20 and the independent occurrence was 5 more than the co-occurrence frequency. In addition, the threshold was set at the 95% confidence ($\chi^2 \geq 3.841$).

- **Manual Cleaning:**
  Based on the χ^2 rules, manual adjustment for the test set was made by only eliminating or adding rules. Additional rules were obtained from the unused "Baseline" rules. The purpose of this experiment was to measure the translation quality when a theoretically perfect cleaning method is applied.

3.5.2 Result of Experiments

The number of transfer rules for each cleaning method is shown in Table 3.3, and the translation quality is shown in Figure 3.6. The number of transfer rules in

Table 3.3. Number of Transfer Rules for Each Cleaning Method

Cleaning Method	Number of Transfer Rules
No Cleaning	92,005
Cutoff by Freq.	10,011
χ^2 Test	922
Manual Cleaning	1,172
(Hand-made Rules)	4,878

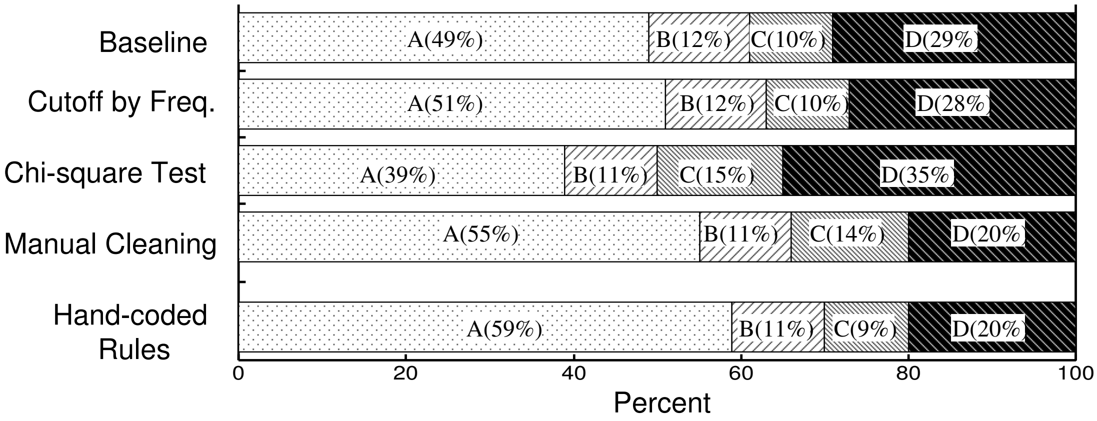


Figure 3.6. Translation Quality for Each Cleaning Method

Table 3.3 is equal to the number of unique pairs of source and target patterns. TDMT integrated with fully hand-coded transfer rules is also shown for reference. The hand-coded rules were created from a different corpus (dialogue corpus; refer to (Furuse et al., 1994)), so it cannot be compared directly, but it contains a sufficient number of appropriate rules.

First, among the fully automatic rule generation methods (Baseline, Cutoff by Frequency, and χ^2 Test), the best method was Cutoff by Frequency, which achieved a 72% translation rate.

In comparison with Baseline, the number of rules decreased to about 1/9 in the case of Cutoff by frequency. However, the translation rate slightly increased. This means that almost all low-frequency rules are redundant or inappropriate, and the Cutoff by Frequency method performs moderately well and is simple.

The χ^2 Test rules are confident from the viewpoint of statistics, but the translation quality was lower. This is because the number of rules was insufficient, and the translations were divided into segments [2]. However, the translation quality did not deteriorate to 1/10 compared with Cutoff by Frequency even though the number of rules decreased to about 1/10. This is because only general rules remained. Therefore, if there are comprehensive and confident rules from the viewpoint of statistics, correct translations can be achieved.

Finally, the translation quality of the Manual Cleaning method was almost the same as that of the Hand-coded Rules. The rules generated from phrase alignment results contained comprehensive rules in the same way as the Hand-made Rules. Therefore, if there is an effective cleaning method, the quality will be close to hand-coded TDMT.

3.5.3 Translation Examples

Table 3.4 shows the translation examples that is translated from an English sentence "*This package is a book, so send it by boat please.*"

"*make te*" is over-generated in the translation number 1 due to an incorrect rule generated from HPA errors. Baseline method often generates additional words because the rule set contains many ungeneralized rules. The translation number 2 is not fluent because the subject is changed from '*package*' to '*hon* (book).' The translation number 3 is not incorrect, but the input sentence is divided into two segments because the number of transfer rules is insufficient. The translation number 4 has no problems.

3.6 Discussion

Corpus Size for Statistical Rule Cleaning The χ^2 test is one of the methods that acquire word translations from a bilingual corpus (Gale and Church, 1991). Since transfer rules are regarded as word correspondences, the hypothesis testing can be applied and correct transfer rules will be acquired. However, a sufficient number of rules could not be acquired (i.e., the coverage was low) in this experiment because of the small corpus.

Melamed (2000) shows an experiment using the Hansard Corpus (English and French). He used 300K bilingual sentences, and extracted translation words with a precision of 87% and coverage of 90%. There were about 41,000 different words for English and 36,000 for French.

Suppose that source and target patterns are regarded in the same way as translation words. 57,000 source patterns and 53,000 target patterns are generated in our experiment. About $\frac{57000*53000}{41000*36000} \simeq 2.0$ times resolution is necessary

[2] TDMT has a partial translation function if there are no rules for parsing.

Table 3.4. Translation Examples from English Sentence “*This package is a book so send it by boat please.*”

No.	Cleaning Method	Rank	Translation
1	Baseline	D	本はこの小包だから船でお願いしまけて *hon wa kono kodutsumi dakara fune de onegai-shi make te.*
2	Cutoff by Freq.	C	本はこの小包だから船で送ってくれますか *hon wa kono kozutsumi dakara fune de okut te kure masu ka.*
3	χ^2 Test	B	この小包は書籍です___お願いします、船で送ります *kono kodutsumi wa shoseki desu ___ onegai-shi masu, fune de okuri masu.*
4	Manual Cleaning	A	この小包は書籍だから船で送ってくれますか *kono kodutsumi wa shoseki dakara fune de okut te kure masu ka.*
–	(Hand-coded)	A	このパッケージが本なので船で送ってください *kono pakkeeji ga hon na no de fune de okut te kudasai.*

in comparison with Melamed (2000)’s experiment, and the number of sentences becomes $300K * 2.0 = 600K$. Consequently, it is estimated that anywhere from a half million to one million bilingual sentences are necessary for statistical rule cleaning.

Longer Sentences The corpus used here contains many short sentences. In the case of long sentences such as newswires, the accuracy of phrase alignment will decrease. However, it can be somewhat maintained if the techniques we described in Section 3.2.2 are applied. In fact, we could expect the problem that the number of transfer rules will increase because longer sentences contain expressions that are more complex. Even though TDMT translates them with short units, a larger corpus will be necessary to maintain coverage of the knowledge.

3.7 Conclusions

Using hierarchical phrase alignment, translation knowledge was acquired from a bilingual corpus of different language families. The acquired knowledge was applied to a translation system, TDMT, and its translation quality was evaluated.

When the transfer rules were cleaned automatically, the translation rate was about 72%.

Hierarchical Phrase Alignment can acquire high coverage rules. If the rules are combined correctly, it is possible to obtain correct translations that are close to hand-coded rules.

Since the corpus contains context-dependent translations and the phrase alignment results have errors, the transfer rules need to be cleaned. Although we used a large corpus of 125K sentences, in which over fifty thousand transfer rules appeared, the rules could not be cleaned to the level that made them as useful and reliable as hand-coded rules.

Future research topics will include enriching our corpus and investigating cleaning methods. The following chapters will describe about some cleaning methods.

Chapter 4

Automatic Construction of Machine Translation Knowledge Using Translation Literalness

4.1 Introduction

Along with the efforts made to accumulate bilingual corpora for many language pairs, quite a few machine translation (MT) systems that automatically construct their knowledge from corpora have been proposed (Brown et al., 1993; Menezes and Richardson, 2001; Imamura, 2002). However, if we use corpora without any restriction, redundant rules are acquired due to translation varieties. Such rules increase ambiguity and may cause inappropriate MT results.

Translation variety increases with corpus size. For instance, large corpora usually contain multiple translations of the same source sentences. Moreover, peculiar translations that depend on context or situation proliferate in large corpora. Our targets are corpora that contain over one hundred thousand sentences.

To reduce the influence of translation variety, we attempt to control the bilingual sentences that are appropriate for machine translation (here called *controlled translation*). Among the measures that can be used for controlled translation, we focus on translation literalness in this chapter. By restricting bilingual sentences during MT knowledge construction, the MT quality will be improved.

The remainder of this chapter is organized as follows. Section 4.2 describes the problems caused by translation varieties. Section 4.3 discusses the kinds of translations that are appropriate for MTs. Section 4.4 introduces the concept of translation literalness and how to measure it. Section 4.5 describes construction methods using literalness, and Section 4.6 evaluates the construction methods.

4.2 Problems Caused by Translation Variety

First, we describe the problems inherent in bilingual corpora when we automatically construct MT knowledge.

4.2.1 Context/Situation-dependent Translation

Some bilingual sentences in corpora depend on the context or situation, and these are not always correct in different contexts.

For instance, the English determiner '*the*' is not generally translated into Japanese. However, when a human translator cannot semantically identify the following noun, a determinant modifier such as '*watashi-no* (my)' or '*sono* (its)' is supplied.

As an example of a situation-dependent translation, the Japanese sentence "*Shashin wo tot-te itadake masu ka?* (Could you take our photograph?)" is sometimes translated into an English sentence as "*Could you press this shutter button?*" This translation is correct from the viewpoint of meaning, but it can only be applied when we want a photograph to be taken. Such examples show that most context/situation-dependent translations are non-literal.

MT knowledge constructed from context/situation-dependent translations cause incorrect target sentences, which may contain omissions or redundant words, when it is applied to an inappropriate context or situation.

4.2.2 Multiple Translations

Generally speaking, a single source expression can be translated into multiple target expressions. Therefore, a corpus contains multiple translations even though they are translated from the same source sentence. For example, the Japanese sentence "*Kono toraberaazu chekku wo genkin ni shite kudasai*" can be translated into English any of the following sentences.

- *I'd like to cash these traveler's checks.* (declarative)
- *Could you change these traveler's checks into cash?* (interrogative)
- *Please cash these traveler's checks.* (imperative)

These translations are all correct. Actually, the corpus of Takezawa et al. (2002) contains ten different translations of this source sentence. When we construct MT knowledge from corpora that contain such variety, redundant rules are acquired. For instance, a transfer-based MT system described in Section 3.3 acquires different transfer rules from each multiple translations, although only one rule is necessary for translating a sentence. Redundant rules increase ambiguity or decrease translation speed (Meyers et al., 2000).

4.3 Appropriate Translation for MTs

4.3.1 Controlled Translation

Controlled language (Mitamura et al., 1991; Mitamura and Nyberg, 1995; Huijsen, 1998) is proposed for monolingual processing in order to reduce variety. This method allows monolingual texts within a restricted vocabulary and a restricted grammar. Texts written by the controlled language method have fewer semantic and syntactic ambiguities when they are read by a human or analyzed by a computer.

A similar idea can be applied to bilingual corpora. Namely, the expressions in bilingual corpora should be restricted, and "translations that are appropriate for the MT" should be used in knowledge construction. This approach assumes that context/situation-dependent translations should be removed before construction so that ambiguities in MT can be decreased. Restricted bilingual sentences are called controlled translations in this thesis.

The following measures are assumed to be available for controlled translation. First three measures are for each of the bilingual sentences in the corpus and the fourth measure is for the whole corpus:

- **Literalness**
  Few omissions or redundant words appear between the source and target sentences. In other words, most words in the source sentence correspond to some words in the target sentence.

- **Context-freeness**
  Source word sequences correspond to the target word sequences independent of the contextual information. With this measure, partial translation can be reused in other sentences.

- **Word-order Agreement**
  The word order of a source sentence agrees substantially with that of a target sentence. This measure ensures that the cost of word re-ordering is small.

- **Word Translation Stability**
  A source word is better translated into the same target word through the corpus.

  For example, the Japanese adjectival verb '*hitsuyoo-da*' can be translated into the English adjective '*necessary*,' the verb '*need*,' or the verb '*require*.' It is better for an MT system to always translate this word into '*necessary*,' if possible.

Effective measures of controlled translation depend on MT methods. For example, word-level statistical MT (Brown et al., 1993) translates a source sentence with a combination of word transfer and word reordering. Thus, word-order agreement is an important measure. On the other hand, this is not important for transfer-based MTs because the word order can be significantly changed through syntactic transfer. A transfer-based MT method using the phrase structure is studied here.

4.3.2 Base MT System

We use Hierarchical Phrase Alignment-based Translator (HPAT) (Imamura, 2002) as the target transfer-based MT system. HPAT is a new version of Transfer Driven Machine Translator (TDMT) (Furuse and Iida, 1994), and their MT engines are the same. However, Transfer rules of HPAT are automatically acquired from a parallel corpus, but those of TDMT were constructed manually. In the rest of this thesis, we call the MT system, which utilizes the transfer rules automatically constructed by the method described in Chapter 3, 'HPAT,' and call the MT system, which utilizes the transfer rules constructed manually, 'TDMT.'

The procedure of HPAT is briefly described as follows (Figure 4.1). First, phrasal correspondences are hierarchically extracted from a parallel corpus using Hierarchical Phrase Alignment (c.f., Chapter 2). Next, the hierarchical correspondences are transferred into patterns, and transfer rules are generated. At the time of translation, the input sentence is parsed by using source patterns in the transfer rules. The MT result is generated by mapping the source patterns to the target patterns. Ambiguities, which occur during parsing or mapping, are solved by selecting the patterns that minimize the semantic distance between the input sentence and the source examples (real examples in the training corpus). Details were described in Chapter 3.

4.3.3 Appropriate Translation for Transfer-based MT

In order to verify effective measures of controlled translation for transfer-based MTs, we review the fundamentals of TDMT in this section.

TDMT was trained by human rule writers. They selected bilingual sentences from a corpus one by one and added or arranged the transfer rules in order to translate the sentences. The target sentences were then rewritten with the aim of minimizing the number of transfer rules. We believe that this way of rewritten translation is appropriate examples for TDMT.

We compared 6,304 bilingual sentences rewritten for an English-to-Japanese

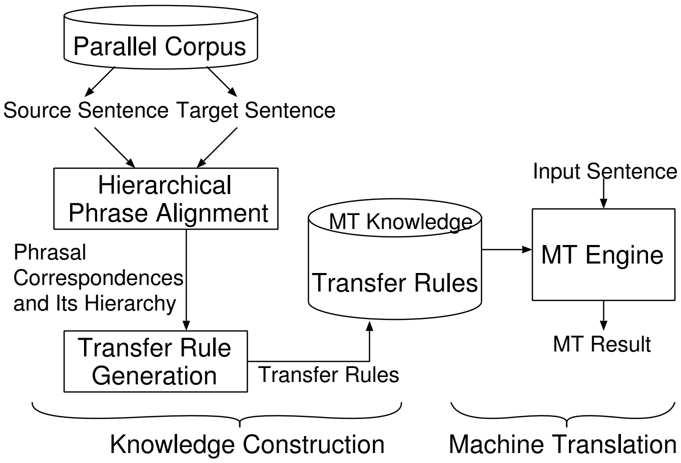


Figure 4.1. Overview of HPAT: Knowledge Construction and Translation Process

version of TDMT and the original translations in the corpus [1]. The statistics in Table 4.1 show that the following measures are effective for transfer-based MT. Note that these data were calculated from the results of morphological analysis and word alignment (c.f., Section 4.6). The correspondences output from the word aligner are called word links.

Literalness Focusing on the number of linked target words in Table 4.1, the value of the rewritten translations is considerably higher than that of the original translations. This result shows that the words of source sentences are translated into target words more directly in the case of the rewritten translations. Thus, the rewritten translations are more literal.

Translation 1 in Table 4.2 is an example of literal translation. "*sono ryokin* (its fare)," which does not appear in the source sentence, is supplied in the original translation, but the words in the rewritten translation correspond to the words in the source sentence with no deficiency and no excess.

Word Translation Stability Focusing on the number of different words in the target language and the mean number of translation words, both values of the rewritten translations are lower than those of the original translations. This is because the rule writers rewrote translations to make target words as simple as

[1] When TDMT translates input sentences already trained, the MT results become identical to the objective translations for the rule writer. Therefore, the rewritten translations were acquired by translating trained sentences by TDMT.

Table 4.1. Comparison of Rewritten Translations and Original Translations

	Rewritten Translations	Original Translations
# of Linked Target Words	28,300 words (49.5%)	20,722 words (34.0%)
# of Different Words in Target Language	3,107 words	3,601 words
Mean # of Translation Words per Source Word	1.51 trans./word	1.94 trans./word
Mean Context-freeness (# of Word Link = 4)	4.45	4.21

Table 4.2. Examples of Rewritten and Original Translations

No.	Type	Sentence
1	Source Sentence	*Are tax and service charges included?*
	Original Translation	その料金は税金とサービス料は込みですか *sono ryokin wa zeikin to saabisuryo wa komi desu ka*
	Rewritten Translation	税とサービス料は含まれていますか *zei to saabisuryo wa fukuma re te i masu ka*
2	Source Sentence	*Is breakfast included?*
	Original Translation	朝食はついていますか *choshoku wa tui te i masu ka*
	Rewritten Translation	朝食は含まれていますか *choshoku wa fukuma re te i masu ka*
3	Source Sentence	*What's the difference between the rate for a single and a twin?*
	Original Translation	料金はシングルとツインではどのくらい違いますか *ryokin wa singuru to tsuin dewa donokurai chigai masu ka*
	Rewritten Translation	シングルとツインの料金の違いはどれくらいですか *singuru to tsuin no ryokin no chigai wa dorekurai desu ka*

possible, and thus the variety of target words was decreased. In other words, the rewritten translations are more stable from the viewpoint of word translation.

For example, focusing on Translations 1 and 2 in Table 4.2, '*include*' is translated into different words in the original translation (Translation 1 is '*komi*', and Translation 2 is '*tsui*'). However, in the rewritten translations, the translated words of '*include*' are stable as '*fukuma*.'

Context-freeness Mean context-freeness in Table 4.1 denotes the mean number of word-link combinations in which word sequences of the source and the target contain word links only between their constituents (cross-links are allowed). If a bilingual sentence can be divided into many translation parts, this value become high. This value depends on the number of word links. When it is calculated only from the sentences that contain four word links, the value of the rewritten translations is slightly higher than that of the original translations.

Translation 3 in Table 4.2 is an example of the context-free translation. A noun phrase "*the rate for a single and a twin*" is locally translated into "*singuru to tsuin no ryokin*" in the rewritten translation. Thus, rules generated from the phrase can be reused to the other translation. However, in the original translation, it is translated into two phrases "*ryokin wa*" and "*singuru to tsuin dewa*," which modify the verb '*chigai*.' Thus, the rules generated from the phrase cannot be reused unless the rule is generated with its modifyee.

4.4 Translation Literalness

We particularly focus on the literalness among the controlled translation measures in order to reduce the incorrect rules that result from context/situation-dependent translations. Word translation stability and context freeness must serve as countermeasures for multiple translations, since they ensure that word translations and structures are steady throughout the corpus. However, the reduction of incorrect translations is done prior to the reduction of ambiguities.

4.4.1 Literalness Measure

A literal translation means that source words are translated one by one to target words. Therefore, a bilingual sentence that has many word correspondences is literal. The word correspondences can be acquired by referring to translation dictionaries or using statistical word aligners (e.g., (Melamed, 2000)).

However, not all source words always have an exact corresponding target word. For example, in the case of English and Japanese, some prepositions are not translated into Japanese. On the contrary, the preposition '*after*' may be translated into Japanese as the noun '*ato*.' These examples show that some functional words have to be translated while others do not. Thus, literalness is not determined only by counting word correspondences but also by estimating how many words in the source and target sentences have to be translated.

Based on the above discussion, the translation literalness of a bilingual sentence is measured by the following procedure. Note that a translation dictionary is utilized in this procedure. The dictionary is automatically constructed by gathering the results of word alignment at this time, though hand-made dictionaries may also be utilized. In this process, we assume that one source word corresponds to one target word.

1. Look up words in the translation dictionary by the source word. T_s denotes the number of source words found in the dictionary entries.

2. Look up words in the dictionary by target words. T_t denotes the number of target words found in the definition parts of the dictionary.

3. If there is an entry that includes both the source and target word, the word pair is regarded as the word link. L denotes the number of word links.

4. Calculate the literalness with the following equation, which we call the Translation Correspondence Rate (TCR) in this thesis.

$$TCR = \frac{2L}{T_s + T_t} \tag{4.1}$$

The TCR denotes the portion of the directly translated words among the words that should be translated. This definition is bi-directional, so omission and redundancy can be measured equally. Moreover, the influence of the dictionary size is low because the words that do not appear in the dictionary are ignored.

For example, suppose that a Japanese source sentence (Source) and its English translations (Targets 1 and 2) are given as shown in Figure 4.2. Target 1 is a literal translation, and Target 2 is a non-literal translation, while the meaning is equivalent. When the circled words are those found in the dictionary, T_s is five, and T_t of Target 1 is also five. There are five word links between Source and Target 1, so the TCR is 1.0 by Equation (4.1).

On the other hand, in the case of Target 2, four words are found in the dictionary ($T_t = 4$), and there are three word links. Thus, the TCR is $\frac{2*3}{5+4} \simeq 0.67$, and Target 1 is judged as more literal than Target 2.

The literalness based on the TCR is judged from a tagged result and a translation dictionary. In other words, 'deep analyses' such as parsing are not necessary.

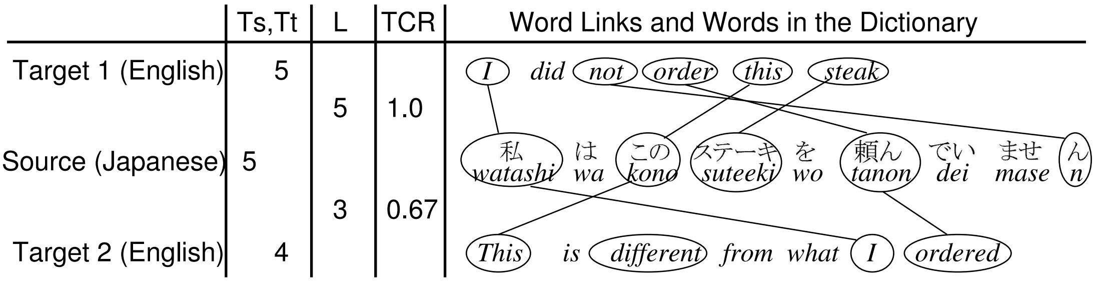

Ts,Tt
L
TCR
Word Links and Words in the Dictionary
Target 1 (English) 5
5 1.0
Source (Japanese) 5
3 0.67
Target 2 (English) 4
I did not order this steak
私 watashi は wa この kono ステーキ suteeki を wo 頼ん tanon でい dei ませ mase ん n
This is different from what I ordered

Figure 4.2. Example of Measuring Literalness Using Translation Correspondence Rate

(Circled words denote words found in the dictionary.
Lines between sentences denote word links.)

4.5 Knowledge Construction Using Translation Literalness

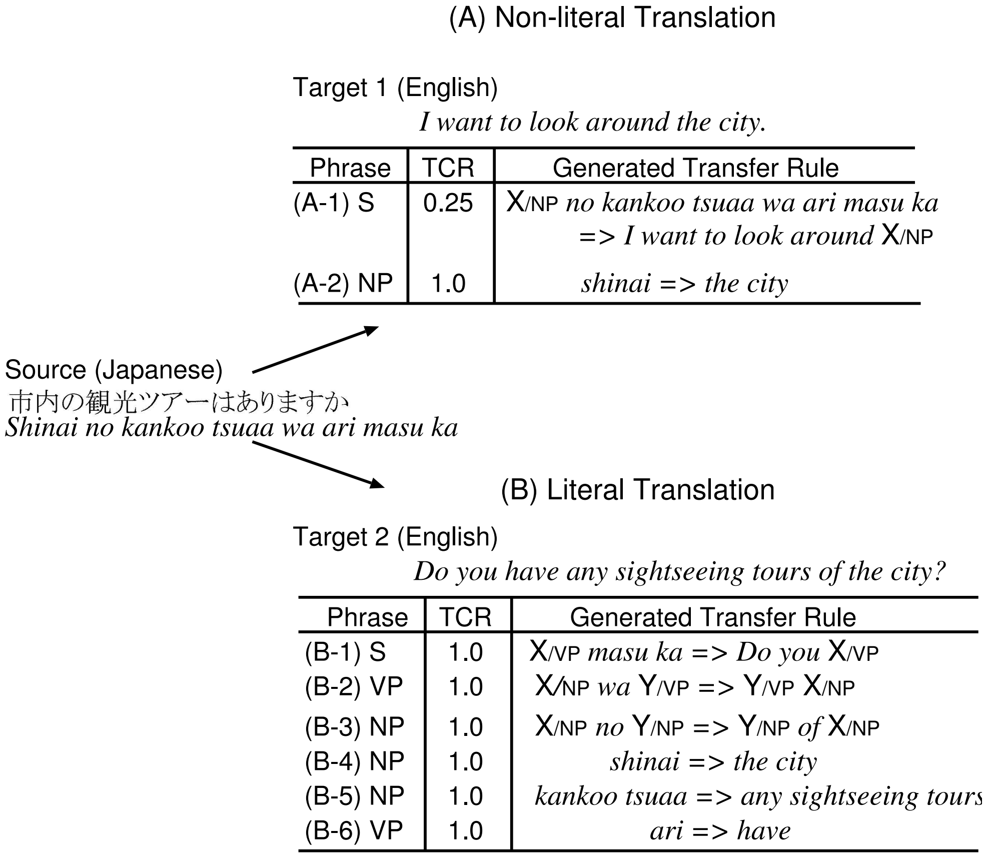
In this section, two approaches for constructing translation knowledge are introduced. One is bilingual corpus filtering, which selects highly literal bilingual sentences from the corpus. Filtering is done as preprocessing before rule acquisition. The other is split construction, which divides a bilingual sentence into literal and non-literal parts and applies different generalization strategies to these parts.

4.5.1 Bilingual Corpus Filtering

We consider two approaches to corpus filtering.

Filtering Based on Threshold A partial corpus is created by selecting bilingual sentences with TCR values higher than a threshold, and MT knowledge is constructed from the extracted corpus. By making the threshold higher, the coverage of MT knowledge will decrease because the size of the extracted corpus becomes smaller.

Filtering Based on Group Maximum First, sentences that have the identical source sentence are grouped together, and a partial corpus is created by selecting the bilingual sentences that have the maximal TCR from each group. As opposed to filtering based on a threshold, all source sentences are used for knowledge construction, so the coverage of MT knowledge can be maintained.



(A) Non-literal Translation

Target 1 (English)
I want to look around the city.

Phrase	TCR	Generated Transfer Rule
(A-1) S	0.25	X/NP *no kankoo tsuaa wa ari masu ka* => *I want to look around* X/NP
(A-2) NP	1.0	*shinai* => *the city*

Source (Japanese)
市内の観光ツアーはありますか
Shinai no kankoo tsuaa wa ari masu ka

(B) Literal Translation

Target 2 (English)
Do you have any sightseeing tours of the city?

Phrase	TCR	Generated Transfer Rule
(B-1) S	1.0	X/VP *masu ka* => *Do you* X/VP
(B-2) VP	1.0	X/NP *wa* Y/VP => Y/VP X/NP
(B-3) NP	1.0	X/NP *no* Y/NP => Y/NP *of* X/NP
(B-4) NP	1.0	*shinai* => *the city*
(B-5) NP	1.0	*kankoo tsuaa* => *any sightseeing tours*
(B-6) VP	1.0	*ari* => *have*

Figure 4.3. Examples of Generated Rules for Japanese-to-English Translation
(A) from Non-literal Translation by Split Construction
(B) from Literal Translation.

However, some context/situation-dependent translations remain in the extracted corpus when only one non-literal translation is in the corpus.

4.5.2 Split Construction into Literal and Other Parts

The TCR can be calculated not only for sentences but also for phrases. In the case of filtering, the coverage of the MT knowledge is decreased by limiting translation to highly literal sentences. However, even though they are non-literal, such sentences may contain literal translations at the phrase level. Thus, the coverage can be maintained if we extract literal phrases from non-literal sentences and construct knowledge from them.

A problem with this approach is that non-literal bilingual sentences sometimes contain idiomatic translations that should not be translated literally. For example, the Japanese greeting "*Hajime mashi te*" should be translated into "*How do you do*," not into its literal translation, "*For the first time*." Such idioms are usually represented by a long word sequence.

To cope with literal and idiomatic translations, a sentence is divided into literal and non-literal parts, and a different construction is applied. Short rules, which are more generalized and easier to reuse, are generated from the literal parts. Long rules, which are stricter in their use in MT, are generated from the non-literal parts. The procedure is described as follows.

1. Phrasal correspondences are acquired by Hierarchical Phrase Alignment.
2. The hierarchy is traced from top to bottom, and the literalness of each correspondence is measured. If the TCR is equal to or higher than the threshold, the phrase is judged as a literal phrase and the tracing stops before reaching the bottom.
3. If the phrase is literal, transfer rules that include its lower hierarchy are generalized.
4. If the top structure (i.e., entire sentence) is not literal, a rule is generated in which only the literal parts are generalized.

For example, suppose that different target sentences from the same source are given as shown in Figure 4.3. The phrase `(A-1)S` has low TCR, but the TCR of the noun phrase pair '*shinai*' and '*the city*' has 1.0. Thus, the phrase `(A-2)NP` is generalized, and the long transfer rule `(A-1)S` is generated from the non-literal translation. On the contrary, the TCR of the top phrase `(B-1)S` is 1.0, so all phrases in (B) are generalized and totally six rules are generated. The rules generated from literal translations are general, and they will be used for the translation of the other sentences.

Thus, by using the split construction, rules like templates are generated from non-literal translations and primary rules for transfer-based MT are generated only from literal phrases. Rules generated from non-literal translations are used only when the input word sequence exactly matches the sequence in the rule. In other words, they are hardly used in different contexts.

4.6 Translation Experiments

In order to evaluate the effect of literalness in MT knowledge construction, we constructed knowledge by using the methods described in Section 4.5 and evaluated the MT quality of the resulting English-to-Japanese translation.

Table 4.3. Statistics of Basic Travel Expression Corpus

Set Name	Feature	English	Japanese
Training	# of Sentences	149,882	
	# of Total Words	868,087	984,197
	# of Different Words	11,288	17,574
	# of Equivalent Phrases	565,208	
	(including Sentence Correspondence)	(619,938)	
Test	# of Sentences	10,150	
	# of Total Words	59,232	67,193
	# of Different Words	4,030	5,040

4.6.1 Experimental Settings

Bilingual Corpus We used the Basic Travel Expression Corpus (BTEC; refer to (Takezawa et al., 2002; Kikui et al., 2003)), which is a collection of Japanese sentences and their English translations based on expressions that are usually found in phrasebooks for foreign tourists. There are many bilingual sentences in which the source sentences are the same but the targets are not. About 13% of different English sentences have multiple Japanese translations. The statistics of BTEC are shown in Table 4.3. We used 149,882 bilingual sentences for the training (i.e., for automatic construction of transfer rules) from BTEC.

Translation Dictionary: Extraction of Word Correspondence For word correspondences that occur more than nine times in the corpus, statistical word alignment was carried out by a similar method to (Melamed, 2000). When words for which the correspondence could not be found remain, a thesaurus (Ohno and Hamanishi, 1984) was used to create correspondences to the words of the same group. The following two translation dictionaries were constructed as a collection of the word correspondences.

- **Dictionary A**
  The dictionary constructed from the results of statistical word alignment and referring to the thesaurus. When word correspondences are extracted by applying this dictionary, the accuracy of word alignment is about 90% for precision and 73% for recall by a closed test of content words. The recall of this dictionary is higher than Dictionary B.

- **Dictionary B**
  The dictionary constructed only from the results of statistical word align-

ment. When word correspondences are extracted by applying this dictionary, the accuracy of word alignment is about 93% for precision and 61% for recall by a closed test of content words.

Evaluation for MT Quality We used the following two metrics to evaluate MT quality.

1. **Automatic Evaluation**

   We used BLUE (Papineni et al., 2002) with 10,150 sentences that were reserved for the test set. The number of references was one for each sentence, and a range from uni-gram to four-gram was used.

2. **Subjective Evaluation**

   From the above-mentioned test set, 510 sentences were evaluated by paired comparison. In detail, the source sentences were translated using the base rule set created from the entire corpus, and the same sources were translated using the rules constructed with literalness. One by one, a Japanese native speaker judged which MT result was better or that they were of the same quality. Subjective quality is represented by the following equation, where I denotes the number of improved sentences and D denotes the number of degraded sentences.

$$\text{Subj. Quality} = \frac{I - D}{\#\text{ of test sentences}} \tag{4.2}$$

4.6.2 Effect of Corpus Filtering Based on Threshold

First, filtering of the training corpus was carried out by changing TCR threshold, transfer rules were constructed from the filtered corpus, and then translation quality was measured. Along with the threshold higher, the training corpus size was decreased. Figures 4.4 and 4.5 show the translation quality according to the number of training sentences (i.e., training corpus size) when Dictionary A and B were applied respectively. The random selection in the graphs denotes that the training corpus size was changed by selecting bilingual sentences randomly.

In the both cases, the translation quality was slightly increased even though the training corpus sizes were decreased according to the threshold of TCR higher. The BLEU scores became maximal when the threshold was $TCR \geq 0.3$, and the scores were 0.233 for Dictionary A and 0.234 for Dictionary B.

When the training corpora were restricted to the higher literal translations, the corpus size was decreased. However, the BLEU scores of the filtering were

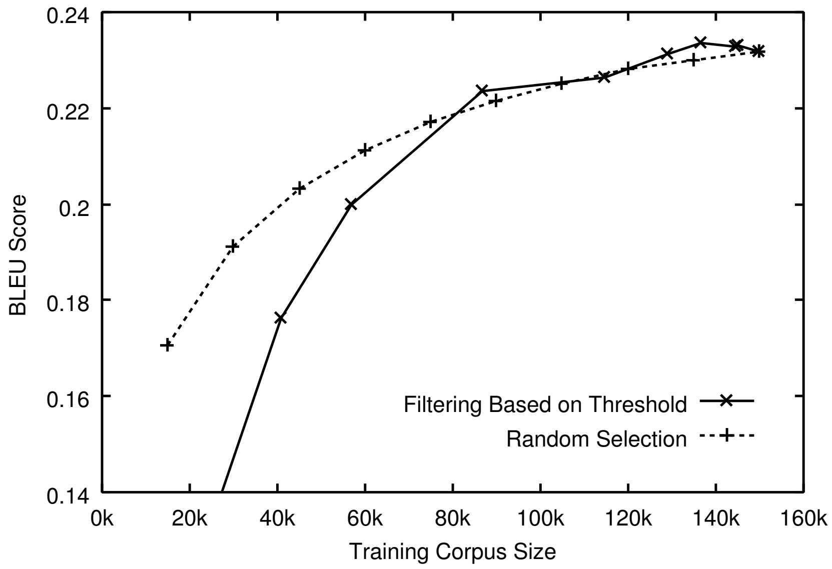


Figure 4.4. Relationship between TCR Threshold for Filtering and Translation Quality (Using Dictionary A)

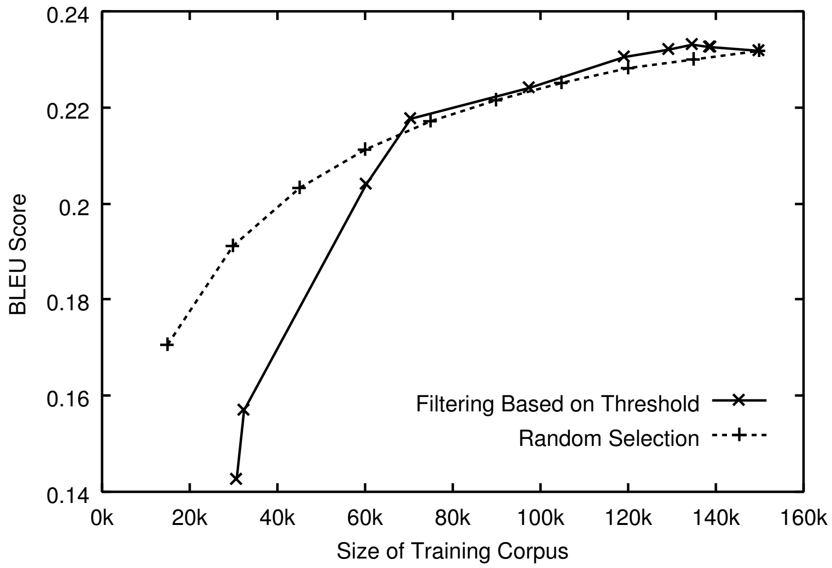


Figure 4.5. Relationship between TCR Threshold for Filtering and Translation Quality (Using Dictionary B)

Table 4.4. MT Quality vs. Construction Methods Using Dictionary B (‘I’ denotes the number of improved sentences, ‘SQ’ denotes the number of the same quality, ‘ST’ denotes the number of the same translation, and ‘D’ denotes the number of degraded sentences.)

Construction Method		# of Translations (Size Ratio)	Coverage of Exact Rules	**BLEU Score**	**Subj. Quality**	
Entire Corpus (Baseline)		149,882 (100%)	65.7%	0.232	—	
Filtering	Threshold ($TCR \geq 0.3$)	134,521 (89.8%)	64.0%	0.233	+3.3%	
					I	30
					SQ	467
					ST	424
					D	13
	Group Maximum	121,623	65.3%	0.240	+1.8%	
					I	30
					SQ	459
					ST	413
					D	21
Split Construction ($TCR \geq 0.8$)		121,623	61.4%	**0.252**	**+8.6%**	
					I	119
					SQ	316
					ST	213
					D	75

higher than those of the random selection while the corpora contain over eighty thousand sentences in the case of Dictionary A and seventy thousand in the case of Dictionary B. When the corpus size was restricted fewer than seventy thousand, the translation quality was extremely degraded. This is because bilingual sentences that are non-literal but are necessary for translation, such as idioms, were removed.

The graphs show us the similar curve, so the coverage of the translation dictionary hardly affects measurement of literalness using TCR.

4.6.3 MT Quality vs. Construction Methods

The level of MT quality achieved by each of the construction methods is compared in Table 4.4. Coverage of exact rules denotes the portion of sentences that were translated by using only the rules that require the source example to exactly match the input sentence. In addition, the threshold $TCR \geq 0.3$ was used for filtering because it was experimentally shown to be the best value. In the case of split construction, we used the extracted corpus after filtering based on the group maximum, and phrases that were $TCR \geq 0.8$ were judged to be literal phrases.

First, focusing on the filtering, the subjective qualities or the BLEU scores are better than the base in both methods. Comparing the threshold with the group maximum, the BLEU score is increased by the group maximum. The coverage of the exact rules is higher even if the corpus size decreases. Filtering based on the group maximum improves the quality while maintaining the coverage of the knowledge.

Although we used a high-density corpus where many English sentences have multiple Japanese translations, the quality improved by only about 2% or 3%. It is difficult to significantly improve the quality by bilingual corpus filtering because it is difficult to both remove insufficiently literal translations and maintain coverage of MT knowledge.

On the other hand, the BLEU score and the subjective quality both improved in the case of split construction, even though the coverage of the exact rules decreased. In particular, the subjective quality improved by about 8.6%. Incorrect translations were suppressed because the rules generated from non-literals are restricted when the MT system applies them.

In summary, all construction methods helped to improve the BLEU scores or the subjective qualities; therefore, construction with translation literalness is an effective way to improve MT quality.

4.7 Related Work

Some machine translation systems automatically rewrite input sentences to the controlled language. For example, the KANT MT system (Mitamura and Nyberg, 2001) checks an input sentence written by humans with the controlled language constraints. If the sentence violates the constraints, the system tries to rewrite it to the controlled language. The system of Shirai et al. (1998) automatically rewrites input sentences to the internal expressions, which are similar to the controlled language, and machine translation is carried out from the internal expressions.

Watanabe et al. (2002) has proposed a method that rewrites target sentences in training corpora to high-frequency expressions using paraphrasing, and constructed a statistical machine translation system. They reported that about 2.5% of translations were improved by applying paraphrasing. If we apply the controlled translation measures to paraphrasing, we can improve not only statistical

MT but also various MT systems with maintaining size of corpora.

4.8 Conclusions

In this chapter, we proposed restricting the translation variety in bilingual corpora by controlled translation, which limits bilingual sentences to the appropriate translations for MT. We focused on literalness from among the various measures for controlled translation and defined a Translation Correspondence Rate for calculating literalness.

Less literal translations could be removed by filtering according to the TCR, and this slightly improved the MT quality.

The TCR is capable of measuring literalness not only for bilingual sentences but also for phrases. In other words, a bilingual sentence can be divided into literal phrases and other phrases. Using this feature, sentences were divided into literal parts and non-literal parts, and transfer rules that could be applied with strong conditions were generated from the non-literal parts. As a result, MT quality as judged by subjective evaluation improved in about 8.6% of the sentences.

Word translation stability and context-freeness were also effective measures. MT quality is expected to be further improved by using these measures because they reduce multiple translations.

Chapter 5

Feedback Cleaning of Machine Translation Rules Using Automatic Evaluation

5.1 Introduction

Along with the efforts made in accumulating bilingual corpora for many language pairs, quite a few machine translation (MT) systems that automatically acquire their knowledge from corpora have been proposed. However, knowledge for transfer-based MT acquired from corpora contains many incorrect/redundant rules due to acquisition errors or translation variety in the corpora. Such rules conflict with other existing rules and cause implausible MT results or increase ambiguity. If incorrect rules could be avoided, MT quality would necessarily improve.

There are two approaches to overcoming incorrect/redundant rules:

- Selecting appropriate rules in a disambiguation process during the translation (on-line processing, (Meyers et al., 2000)).
- Cleaning incorrect/redundant rules after automatic acquisition (off-line processing, (Menezes and Richardson, 2001; Imamura, 2002)).

We employ the second approach in this chapter. The cutoff by frequency (Menezes and Richardson, 2001) and the hypothesis test (Imamura, 2002) have been applied to clean the rules. The cutoff by frequency can slightly improve MT quality, but the improvement is still insufficient from the viewpoint of the large number of redundant rules. The hypothesis test requires very large corpora in order to obtain a sufficient number of rules that are statistically confident.

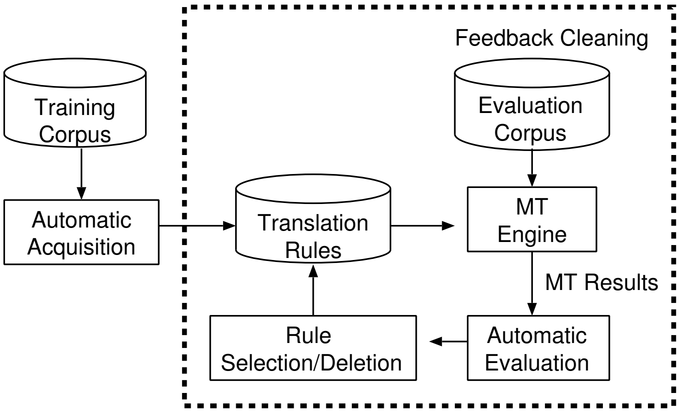


Figure 5.1. Structure of Feedback Cleaning

Another current topic of machine translation is automatic evaluation of MT quality (Papineni et al., 2002; Doddington, 2002; Yasuda et al., 2001; Akiba et al., 2001). These metrics aim to replace subjective evaluation in order to speed up the development cycle of MT systems. However, they can be utilized not only as developers' aids but also for automatic tuning of MT systems (Su et al., 1992).

We propose *feedback cleaning* that utilizes an automatic evaluation for removing incorrect/redundant translation rules as a tuning method (Figure 5.1). Our method evaluates the contribution of each rule to the MT results and removes inappropriate rules as a way to increase the evaluation scores. Since the automatic evaluation correlates with a subjective evaluation, MT quality will improve after cleaning.

Our method only evaluates MT results and does not consider various conditions of the MT engine, such as parameters, interference in dictionaries, disambiguation methods, and so on. Even if an MT engine avoids incorrect/redundant rules by on-line processing, errors inevitably remain. Our method cleans the rules in advance by only focusing on the remaining errors. Thus, our method complements on-line processing and adapts translation rules to the given conditions of the MT engine.

Rule No.	Syn. Cat.	Source & Target Patterns	Source Example
1	VP	$X_{\texttt{VP}}$ *at* $Y_{\texttt{NP}}$ ⇒ Y' *de* X'	((*present*, *conference*) ...)
2	VP	$X_{\texttt{VP}}$ *at* $Y_{\texttt{NP}}$ ⇒ Y' *ni* X'	((*stay*, *hotel*), (*arrive*, *p.m*) ...)
3	VP	$X_{\texttt{VP}}$ *at* $Y_{\texttt{NP}}$ ⇒ Y' *wo* X'	((*look*, *it*) ...)
4	NP	$X_{\texttt{NP}}$ *at* $Y_{\texttt{NP}}$ ⇒ Y' *no* X'	((*man*, *front desk*) ...)

Figure 5.2. Example of HPAT Transfer Rules

5.2 MT System and Problems of Automatic Acquisition

5.2.1 MT Engine

We use the Hierarchical Phrase Alignment-based Translator (HPAT) (Imamura, 2002) as a transfer-based MT system. The detail of HPAT was described in Chapter 3.

The most important knowledge in HPAT is transfer rules, which define the correspondences between source and target language expressions. An example of English-to-Japanese transfer rules is shown in Figure 5.2. The transfer rules are regarded as a synchronized context-free grammar. Symbols like $X_{\texttt{VP}}$ or X' in Figure 5.2 denote corresponding non-terminal symbols.

When the system translates an input sentence, the sentence is first parsed by using source patterns of the transfer rules. Next, a tree structure of the target language is generated by mapping the source patterns to the corresponding target patterns. When non-terminal symbols remain in the target tree, target words are inserted by referring to a translation dictionary.

Ambiguities, which occur during parsing or mapping, are resolved by selecting the rules that minimize the semantic distance between the input words and source examples (real examples in the training corpus) of the transfer rules (Furuse and Iida, 1994). For instance, when the input phrase "*leave at 11 a.m.*" is translated into Japanese, Rule 2 in Figure 5.2 is selected because the semantic distance from the source example (*arrive*, *p.m.*) is the nearest to the head words of the input phrase (*leave*, *a.m.*). The output sentence "*11 ji* (11 a.m.) *ni deru* (leave)" is generated as a result.

5.2.2 Problems of Automatic Acquisition

HPAT automatically acquires its transfer rules from sentence-aligned (parallel) corpora by using Hierarchical Phrase Alignment (c.f., Chapter 2). However, the rule set contains many incorrect/redundant rules. The reasons for this problem are roughly classified as follows, and MT systems, which utilize rules acquired from corpora, cannot avoid these problems

- Errors in automatic rule acquisition
- Translation variety in corpora
    - The acquisition process cannot generalize the rules because bilingual sentences depend on the context or the situation.
    - Corpora contain multiple (paraphrasable) translations of the same source expression.

In the experiments of Chapter 3, about 92,000 transfer rules were acquired from about 120,000 bilingual sentences. Most of these rules are low-frequency. By the experiments, MT quality slightly improved even though the low-frequency rules were removed to a level of about 1/9 the previous number. However, since some of them, such as idiomatic rules, are necessary for translation, MT quality cannot be dramatically improved by only removing low-frequency rules.

5.3 Automatic Evaluation of MT Quality

We utilize BLEU (Papineni et al., 2002) for the automatic evaluation of MT quality in this chapter.

BLEU measures the similarity between MT results and translation results made by humans (called *references*). This similarity is measured by n-gram precision scores. Several kinds of n-grams can be used in BLEU. We use from 1-gram to 4-gram in this chapter, where an 1-gram precision score indicates the adequacy of word translation and longer n-gram (e.g., 4-gram) precision scores indicate fluency of sentence translation. The BLEU score is calculated by the geometric mean of n-gram precision scores, so this measure combines adequacy and fluency.

Note that a sizeable set of MT results is necessary in order to calculate an accurate BLEU score. Although it is possible to calculate the BLEU score of a single MT result, it contains errors from the subjective evaluation. BLEU cancels out individual errors by summing the similarities of MT results. Therefore, we need all of the MT results from the evaluation corpus in order to calculate an accurate BLEU score.

One feature of BLEU is its use of multiple references for a single source sentence. However, one reference per sentence is used in this chapter because an already existing bilingual corpus is applied to the cleaning. In addition, when MT results are Japanese, we have to segment them into words. In this chapter, both MT results and the references are segmented by one morphological analyzer (Yamamoto et al., 1997). Results of the morphological analysis contain errors. However, the influence of the errors on the BLEU score will be restrained because the same sentences share the same errors.

5.4 Feedback Cleaning

In this section, we introduce the proposed method, called feedback cleaning. This method is carried out by selecting or removing translation rules to increase the BLEU score of the evaluation corpus (Figure 5.1). Thus, this task is regarded as a combinatorial optimization problem of translation rules. The hill-climbing algorithm, which involves the features of this task, is applied to the optimization. The following sections describe the reasons for using this method and its procedure. The hill-climbing algorithm often falls into locally optimal solutions. However, we believe that a locally optimal solution is more effective in improving MT quality than the previous methods.

5.4.1 Costs of Combinatorial Optimization

Most combinatorial optimization methods iterate changes in the combination and the evaluation. In the machine translation task, the evaluation process requires the longest time. For example, in order to calculate the BLEU score of a combination (solution), we have to translate C times, where C denotes the size of the evaluation corpus. Furthermore, in order to find the nearest neighbor solution, we have to calculate all BLEU scores of the neighborhood. If the number of rules is R and the neighborhood is regarded as consisting of combinations made by changing only one rule, we have to translate $C \times R$ times to find the nearest neighbor solution. Assume that $C = 10,000$ and $R = 100,000$, the number of sentence translations (sentences to be translated) becomes one billion. It is infeasible to search for the optimal solution without reducing the number of sentence translations.

A feature of this task is that removing rules is easier than adding rules. The rules used for translating a sentence can be identified during the translation. Conversely, the source sentence set $S[r]$, where a rule r is used for the translation, is determined once the evaluation corpus is translated. When r is removed, only the MT results of $S[r]$ will change, so we do not need to re-translate other sentences. In other words, when r is removed, the BLEU score can be calculated by translating only $S[r]$ sentences. On the contrary, to add a rule, the entire corpus must be re-translated because it is unknown which MT results will change by adding a rule.

5.4.2 Cleaning Procedure

Based on the above discussion, we utilize the hill-climbing algorithm, in which the initial solution contains all rules (called the base rule set) and the search for a combination is done by only removing rules. The algorithm is shown in Figure 5.3. This algorithm can be summarized as follows.

1. Translate the evaluation corpus first and then obtain the rules used for the translation and the BLEU score before removing rules.

2. For each rule one-by-one, calculate the BLEU score after removing the rule and obtain the difference between this score and the score before the rule was removed. This difference is called the *rule contribution*.

3. If the rule contribution is negative (i.e., the BLEU score increases after removing the rule), remove the rule.

4. Repeat Steps 1 to 3 until BLEU score cannot be improved.

In order to achieve faster convergence, this algorithm removes all rules whose rule contribution is negative in one iteration. This assumes that the removed rules are independent from one another.

Assuming that five rules on average are applied to translate a sentence, the number of sentence translations becomes $C + 5 \times C = 60,000$ for one iteration (testing all rules).

5.5 N-fold Cross-cleaning

In general, most evaluation corpora are smaller than training corpora. Therefore, omissions of cleaning will remain because not all rules can be tested by the evaluation corpus. In order to avoid this problem, we propose an advanced method called *cross-cleaning* (Figure 5.4), which is similar to cross-validation.

The procedure of cross-cleaning is as follows.

1. First, create the base rule set from the entire training corpus.

```
static: $C_{eval}$, an evaluation corpus
        $R_{base}$, a rule set acquired from the entire training corpus (the base rule set)
        $R$, a current rule set, a subset of the base rule set
        $S[r]$, a source sentence set where the rule $r$ is used for the translation
        $Doc_{iter}$, an MT result set of the evaluation corpus
        translated with the current rule set

procedure CLEAN-RULESET ()
    $R \leftarrow R_{base}$
    repeat
        $R_{iter} \leftarrow R$
        $R_{remove} \leftarrow \emptyset$
        $score_{iter} \leftarrow$ SET-TRANSLATION()
        for each $r$ in $R_{iter}$ do
            if $S[r] \neq \emptyset$ then
                $R \leftarrow R_{iter} - \{r\}$
                translate all sentences in $S[r]$, and obtain the MT results $T[r]$
                $Doc[r] \leftarrow$ the MT result set that $T[r]$ is replaced from $Doc_{iter}$
                the rule contribution $contrib[r] \leftarrow score_{iter} -$ BLEU-SCORE($Doc[r]$)
                if $contrib[r] < 0$ then add $r$ to $R_{remove}$
        end
        $R \leftarrow R_{iter} - R_{remove}$
    until $R_{remove} = \emptyset$

function SET-TRANSLATION () returns a BLEU score of the evaluation corpus
translated with $R$
    $Doc_{iter} \leftarrow \emptyset$
    for each $r$ in $R_{base}$ do $S[r] \leftarrow \emptyset$ end
    for each $s$ in $C_{eval}$ do
        translate $s$ and obtain the MT result $t$
        obtain the rule set $R[s]$ that is used for translating $s$
        for each $r$ in $R[s]$ do add $s$ to $S[r]$ end
        add $t$ to $Doc_{iter}$
    end
    return BLEU-SCORE($Doc_{iter}$)
```

Figure 5.3. Feedback Cleaning Algorithm

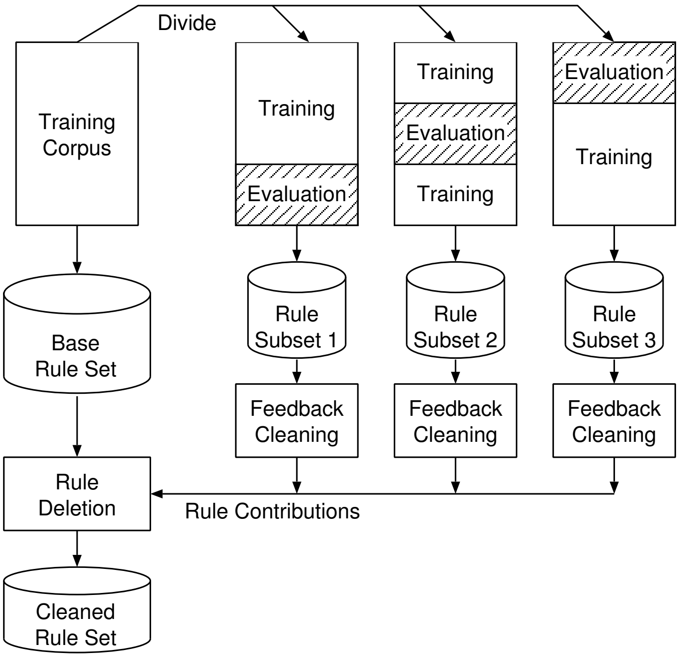


Figure 5.4. Structure of Cross-cleaning
(In the case of three-fold cross-cleaning)

2. Next, divide the training corpus into N pieces uniformly.
3. Leave one piece for the evaluation, acquire rules from the rest $(N-1)$ of the pieces, and repeat them N times. Thus, we obtain N pairs of rule set and evaluation sub-corpus. Each rule set is a subset of the base rule set.
4. Apply the feedback cleaning algorithm to each of the N pairs and record the rule contributions even if the rules are removed. The purpose of this step is to obtain the rule contributions.
5. For each rule in the base rule set, sum up the rule contributions obtained from the rule subsets. If the sum is negative, remove the rule from the base rule set.

The major difference of this method from cross-validation is Step 5. In the case of cross-cleaning, the rule subsets cannot be directly merged because some rules have already been removed in Step 4. Therefore, we only obtain the rule contributions from the rule subsets and sum them up. The summed contribution is an approximate value of the rule contribution to the entire training corpus.

Table 5.1. Statistics of Basic Travel Expression Corpus

Set Name	Feature	English	Japanese
Training Corpus	# of Sentences	149,882	
	# of Total Words	868,087	984,197
	# of Different Words	11,288	17,574
Evaluation Corpus	# of Sentences	10,145	
	# of Total Words	59,533	67,544
	# of Different Words	4,013	4,986
Test Corpus	# of Sentences	10,150	
	# of Total Words	59,232	67,193
	# of Different Words	4,030	5,040

Cross-cleaning removes the rules from the base rule set based on this approximate contribution.

Cross-cleaning uses all sentences in the training corpus, so it is nearly equivalent to applying a large evaluation corpus to feedback cleaning, even though it does not require specific evaluation corpora.

5.6 Evaluation

In this section, the effects of feedback cleaning are evaluated by using English-to-Japanese translation. First, in Section 5.6.1, we describe experimental settings. In Section 5.6.2, we carry out feedback cleaning using an evaluation corpus in order to observe the characteristics of feedback cleaning. In Section 5.6.3, we compare MT quality of previous and proposed methods, including cross-cleaning. Finally, in Section 5.6.4, we show examples of removed rules by feedback cleaning.

5.6.1 Experimental Settings

Bilingual Corpora The corpus used in the following experiments is the Basic Travel Expression Corpus (Takezawa et al., 2002; Kikui et al., 2003). This is a collection of Japanese sentences and their English translations based on expressions that are usually found in phrasebooks for foreign tourists. We divided it into sub-corpora for training, evaluation, and test as shown in Table 5.1. The number of rules acquired from the training corpus (the base rule set size) was 105,588.

Evaluation Methods of MT Quality We used the following two methods to evaluate MT quality.

1. **Test Corpus BLEU Score**
   The BLEU score was calculated with the test corpus. The number of references was one for each sentence, in the same way used for the feedback cleaning.

2. **Subjective Quality**
   A total of 510 sentences from the test corpus were evaluated by paired comparison. Specifically, the source sentences were translated using the base rule set, and the same sources were translated using the rules after the cleaning. One-by-one, a Japanese native speaker judged which MT result was better or that they were of the same quality. Subjective quality is represented by the following equation, where I denotes the number of improved sentences and D denotes the number of degraded sentences.

$$\text{Subj. Quality} = \frac{I - D}{\#\text{ of test sentences}} \tag{5.1}$$

5.6.2 Feedback Cleaning Using Evaluation Corpus

In order to observe the characteristics of feedback cleaning, cleaning of the base rule set was carried out by using the evaluation corpus. The results are shown in Figure 5.5. This graph shows changes in the test corpus BLEU score, the evaluation corpus BLEU score, and the number of rules along with the number of iterations.

Consequently, the removed rules converged at nine iterations, and 6,220 rules were removed. The evaluation corpus BLEU score was improved by increasing the number of iterations, demonstrating that the combinatorial optimization by the hill-climbing algorithm worked effectively. The test corpus BLEU score reached a peak score of 0.245 at the second iteration and slightly decreased after the third iteration due to overfitting. However, the final score was 0.244, which is almost the same as the peak score.

The test corpus BLEU score was lower than the evaluation corpus BLEU score because the rules used in the test corpus were not exhaustively checked by the evaluation corpus. If the evaluation corpus size could be expanded, the test corpus score would improve.

About 37,000 sentences were translated on average in each iteration. This means that the time for an iteration is estimated at about ten hours if translation

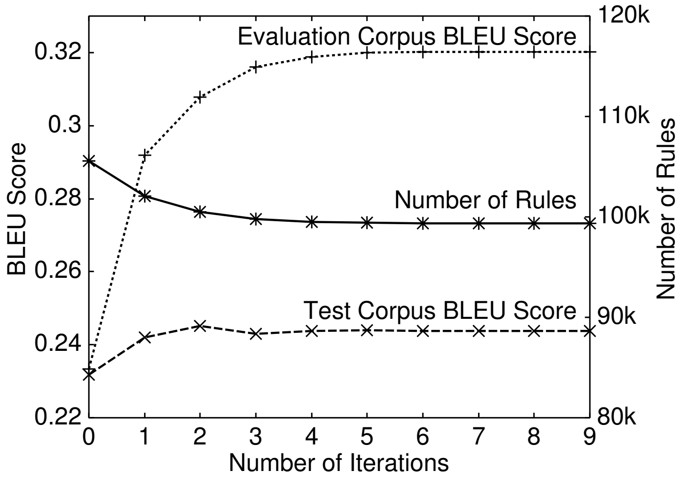


Figure 5.5. Relationship between Number of Iterations and BLEU Scores/Number of Rules

speed is one second per sentence. This is acceptable time for us because our method does not require real-time processing. [1]

5.6.3 MT Quality vs. Cleaning Methods

Next, in order to compare the proposed methods with the previous methods, the MT quality achieved by each of the following five methods was measured.

1. **Baseline**
   The MT results using the base rule set.

2. **Cutoff by Frequency**
   Low-frequency rules that appeared in the training corpus less often than twice were removed from the base rule set. This threshold was experimentally determined by the test corpus BLEU score.

3. χ^2 **Test**
   The χ^2 test was performed in the same manner as in the experiment of Imamura (2002). We introduced rules with more than 95 percent confidence ($\chi^2 \geq 3.841$).

[1] In this experiment, it took about 80 hours until convergence using a Pentium 4 2-GHz computer.

Table 5.2. MT Quality vs. Cleaning Methods
(I denotes the number of improved sentences, SQ denotes the number of the same quality, ST denotes the number of the same translation, and D denotes the number of degraded sentences.)

Cleaning Method		# of Rules	**Test Corpus BLEU Score**	**Subj. Quality**	
Baseline		105,588	0.232	—	
Previous Methods	Cutoff by Freq.	26,053	0.234	+2.35%	
				I	87
				SQ	348
				ST	(257)
				D	75
	χ^2 Test	1,499	0.157	-5.88%	
				I	119
				SQ	242
				ST	(114)
				D	149
Proposed Methods	Simple FC	99,368	**0.244**	**+5.69%**	
				I	79
				SQ	381
				ST	(266)
				D	50
	Cross-cleaning	82,462	**0.277**	**+11.18%**	
				I	107
				SQ	353
				ST	(234)
				D	50

4. **Simple Feedback Cleaning**
   Feedback cleaning was carried out using the evaluation corpus in Table 5.1. This is the same experiment in Section 5.6.2.

5. **Cross-cleaning**
   N-fold cross-cleaning was carried out. We applied five-fold cross-cleaning in this experiment.

The results are shown in Table 5.2. This table shows that the test corpus BLEU score and the subjective quality of the proposed methods (simple feedback cleaning and cross-cleaning) are considerably improved over those of the previous methods.

Focusing on the subjective quality of the proposed methods, some MT results were degraded from the baseline due to the removal of rules. However, the subjective quality levels were relatively improved because our methods aim to increase the portion of correct MT results.

Focusing on the number of the rules, the rule set of the simple feedback cleaning is clearly a locally optimal solution, since the number of rules is more than that of cross-cleaning, although the BLEU score is lower. In comparing the number of rules in cross-cleaning with that in the cutoff by frequency, the former is three times higher than the latter. We assume that the solution of cross-cleaning is also the locally optimal solution. If we could find the globally optimal solution, the MT quality would certainly improve further.

5.6.4 Examples of Removed Rules

Figure 5.6 shows the examples of removed rules by the feedback cleaning using the evaluation corpus and translations changed before and after removing the rules. They have the following features.

- Rule 1 in Figure 5.6 is a wrong rule that '*admission*' has never translated to the target language. Such wrong rules cause incorrect MT results that lack important constituents. However, the feedback cleaning removes the rules because similarity between MT results and the references decreases.

- Rule 2 translates the English verb phrase "*include tax*" into the Japanese predicate phrase "zei komi da." This rule is not wrong. However, the rule is removed when the final MT results become incorrect by combining with other correct rules.

- Rule 3 is correct, and both translations before and after removing the rule are correct. when correct rules conflict with each other, minor rules, which are rarely used in the evaluation corpus, are removed like this example.

- Rule 4 is correct, but it was removed as well as Rule 3. Then, a few translations after removal became incorrect. The MT engine used here selects appropriate rules based on the semantic distance between the input sentence and the source examples. Rules used after removal depend on the input, so some translations after removal may become incorrect. However, as we described in Section 5.2, the feedback cleaning removes rules in order to increase the portion of correct MT results.

No.	Transfer Rule and Translation Examples of Evaluation Corpus		
	Syn. Cat.	Source & Target Patterns	Source Example
1	NP	*the admission* $\mathrm{X}_{\texttt{N}} \Rightarrow$ X'	((*fee*) ...)
	Input	*What is the admission fee?*	
	Translation before Removal	料金はいくらですか. *ryokin wa ikura desu ka*	
	Translation after Removal	入場料はいくらですか. *nyujoryo wa ikura desu ka*	
2	VP	*include* $\mathrm{X}_{\texttt{NP}} \Rightarrow$ X' *komi da*	((*tax*) (*gas*) ...)
	Input	*Does it include tax?*	
	Translation before Removal	税込みれていますか. *zei komi re te i masu ka*	
	Translation after Removal	税金は含まれていますか. *zeikin wa fukuma re te i masu ka*	
3	S	*please* $\mathrm{X}_{\texttt{VP}} \Rightarrow$ X' *tai no desu ga*	((*send*) (*receive*) ...)
	Input	*Please cash this traveller's check.*	
	Translation before Removal	このトラベラーズチェックを現金にしたいのですが. *kono traberaazu chekku wo genkin ni shi tai no desu ga*	
	Translation after Removal	このトラベラーズチェックを現金にしてください. *kono traberaazu chekku wo genkin ni shi te kudasai*	
4	S	*could you* $\mathrm{X}_{\texttt{VP}} \Rightarrow$ X' *te kudasai*	((*check*) (*find*) ...)
	Input	*Could you tell me how to fill in this form?*	
	Translation before Removal	この書類の書き方を教えてください. *kono shorui no kakikata wo oshie te kudasai*	
	Translation after Removal	この書類の書き方を教えていただけませんか. *kono shorui no kakikata wo oshie te itadake mase n ka*	
	Input	*Could you give me a hand for a second?*	
	Translation before Removal	ちょっと手伝ってください. *chotto tetsudat te kudasai*	
	Translation after Removal	ちょっと手伝います. *chotto tetsudai masu*	

Figure 5.6. Example of Removed Rules by Feedback Cleaning Using Evaluation Corpus

5.7 Discussion

5.7.1 Other Automatic Evaluation Methods

The idea of feedback cleaning is independent of BLEU. Some automatic evaluation methods of MT quality other than BLEU have been proposed. For example, Su et al. (1992), Yasuda et al. (2001), and Akiba et al. (2001) measure similarity between MT results and the references by DP matching (edit distances) and then output the evaluation scores. Doddington (2002) measures similarity based on n-gram precision as well as BLEU. These automatic evaluation methods that output scores are applicable to feedback cleaning.

The characteristics common to these methods, including BLEU, is that the similarity to references are measured for each sentence, and the evaluation score of an MT system is calculated by aggregating the similarities. Therefore, MT results of the evaluation corpus are necessary to evaluate the system, and reducing the number of sentence translations is an important technique for all of these methods.

The effects of feedback cleaning depend on the characteristics of objective measures. DP-based measures and BLEU have different characteristics (Yasuda et al., 2003). The exploration of several measures for feedback cleaning remains an interesting future work.

5.7.2 Optimization for Other Machine Translation Methods

Statistical machine translation (Brown et al., 1993; Och and Ney, 2002) is another approach to corpus-based machine translation. Och (2003) proposed an optimization method for statistical MT by using automatic evaluation. This method optimizes weights of feature functions, which are utilized for learning statistical models, for maximizing automatic evaluation scores.

Contrasting our proposed methods with Och (2003)'s one, the concepts that aim to improve MT quality by maximizing the automatic evaluation scores are the same. However, variables that will be optimized are quite different. Och (2003) optimized eight variables represented by real number. Our proposed methods optimize the combination of about 100,000 rules, which is regarded as 100,000 variables represented by two-value. If Och (2003)'s method was applied to the cleaning of the rules, a solution would not be searched for because of the huge number of variables. If our methods were applied to the optimization of statistical MT, the variables would be changed to the real numbers. Therefore, we conclude that these methods are not compatible with each other.

However, our methods can be applied to any MT methods that utilize translation rules. For instance, example-based MT (Nagao, 1984) translates an input sentence by retrieving bilingual sentences (or partial sentences, called examples) that are similar to the input sentence from corpora, and modifying a target part of the examples. Even though example-based MT selects the best example by referring to thesauri when multiple examples are retrieved, it occasionally selects an incorrect example. In this case, our methods can remove inappropriate examples that cause incorrect MT results.

In addition, when we manually construct MT rules (e.g., (Ikehara et al., 1991)), additional rules frequently impede the application of existing rules (i.e., side effects occur) as the rule set grows larger. Even in this case, the side effects are restrained by measuring rule contribution of the additional rules.

5.7.3 Domain Adaptation

When applying corpus-based machine translation to a different domain, bilingual corpora of the new domain are necessary. However, the sizes of the new corpora are generally smaller than that of the original corpus because the collection of bilingual sentences requires a high cost.

The feedback cleaning proposed in this chapter can be interpreted as adapting the translation rules so that the MT results become similar to the evaluation corpus. Therefore, if we regard the bilingual corpus of the new domain as the evaluation corpus and carry out feedback cleaning, the rule set will be adapted to the new domain. In other words, our method can be applied to adaptation of an MT system by using a smaller corpus of the new domain (Paul et al., 2003).

5.8 Conclusions

In this chapter, we proposed a feedback cleaning method that utilizes automatic evaluation to remove incorrect/redundant translation rules. BLEU was utilized for the automatic evaluation of MT quality, and the hill-climbing algorithm was applied to searching for the combinatorial optimization. Utilizing features of this task, incorrect/redundant rules were removed from the initial solution, which contains all rules acquired from the training corpus. In addition, we proposed n-fold cross-cleaning to reduce the influence of the evaluation corpus size. Our experiments show that the MT quality was improved by 11% in paired comparison and by 0.045 in the BLEU score. This is considerable improvement over the previous methods.

Chapter 6

Conclusions

6.1 Summary

In this thesis, we proposed construction methods of machine translation knowledge from bilingual corpora (Figure 6.1).

Chapter 1 gave an overview of corpus-based machine translation and the objective of this thesis. Corpus-based MT is the opposite concept of hand-coded MT, and it is roughly classified into two approaches: example-based MT and statistical MT. The author's target is automatic construction of translation knowledge for example-based MTs. Some automatic acquisition methods for example-based MTs have already been proposed. However, few methods have been evaluated through use with actual machine translation systems. The goal of this thesis was not only to propose an acquisition method but also to apply the acquired knowledge to an MT system and to identify and solve the inherent problems of corpus-based MT.

In Chapter 2, the author proposed a *hierarchical phrase alignment* (HPA) method, which is at the heart of this work's automatic construction of MT knowledge. HPA automatically extracts equivalent phrases, which are corresponding expressions between bilingual sentences, and it employs parsers. Although previous methods extract correspondences after determining the parsing trees of the bilingual sentence, this method simultaneously extracts the best parsing trees and corresponding phrases by utilizing a similarity measure called the *phrase correspondence score*. This approach has two features: 1) some ambiguities of parsing are resolved by using structural similarity of the bilingual sentence, and 2) if the parsing fails, HPA outputs the sequence of partial trees and correspondences. The best sequence is quickly searched for by using the forward DP backward A^* algorithm. Using this method, about twice as many equivalent phrases were extracted than by the previous methods, and almost no deterioration was observed.

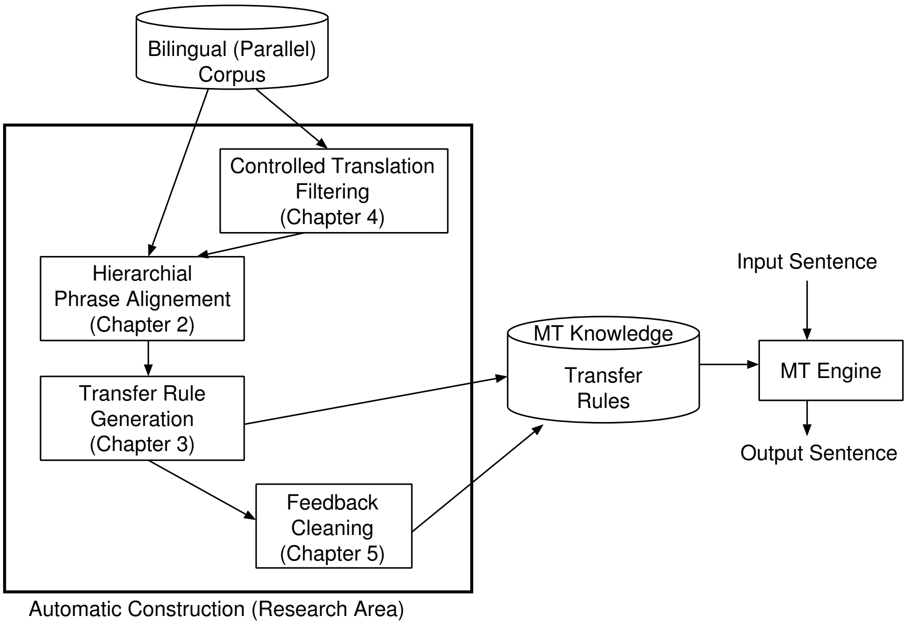

Bilingual (Parallel) Corpus
Controlled Translation Filtering (Chapter 4)
Hierarchial Phrase Alignement (Chapter 2)
Transfer Rule Generation (Chapter 3)
Feedback Cleaning (Chapter 5)
Automatic Construction (Research Area)
MT Knowledge
Transfer Rules
Input Sentence
MT Engine
Output Sentence

Figure 6.1. Structure of Thesis (Duplication of Figure 1.2)

Table 6.1. Causes of Incorrect/Redundant Rules (Duplication of Table 3.1)

	(1) Incorrect Translation	(2) Ambiguity
(a) Problems in Corpora	(1-a) Context/Situation Dependent Translation	(2-a) Multiple Expressions
(b) Problems of HPA	(1-b) Incorrect Phrase Alignment	(2-b) Lack of Correspondence

In Chapter 3, the HPA method proposed in Chapter 2 was applied to a large corpus, and the translation knowledge was automatically constructed. The knowledge was integrated into the MT engine, which is based on transfer driven machine translation (TDMT). Then, translation quality was measured. Through this integration, the problem of the knowledge containing many redundant rules became clear. The reasons for this are classified in Table 6.1. These problems are fundamentally inherent in corpus-based machine translation. By eliminating the redundant rules, quality of MT that automatically constructed from bilingual corpora became close to a hand-coded MT.

Chapter 4 focused on the problems caused by bilingual corpora themselves ((a) in Table 6.1). Bilingual corpora contain not only sentences that are appropriate for MT but also sentences that are inappropriate due to translation variety. For example, bilingual corpora usually contain context/situation-dependent translations or multiple translations even though the source sentences are equal. If we construct translation knowledge from such corpora, many redundant rules are generated, and these cause incorrect MT results or increase ambiguity. In the first half of this chapter, we discussed the kinds of bilingual sentences that are appropriate for MT (called *controlled translation*). Consequently, it was found that literalness, word translation stability, and context-freeness are effective for a syntactic transfer method.

The latter half of Chapter 4 focused on *literalness* among the above controlled translation metrics and defined the *translation correspondence rate* (TCR) as a measure of literalness. Based on TCR, two knowledge construction methods were proposed. One is to filter the corpus, which collects high literal (i.e., high TCR) bilingual sentences, before knowledge acquisition. This method could not dramatically improve MT quality because it removed necessary translations for MTs, such as idiomatic expressions. The other is the split-construction method, which divides a bilingual sentence into literal parts and the other parts before different generalizations are applied. This method is based on the assumption that TCR can capture literalness not only in sentences but also in phrases. By using split construction, MT quality was improved by about 8.6% measured by paired comparison.

Chapter 5 discussed the post-processing done in the automatic acquisition of MT knowledge. The knowledge acquired from bilingual corpora contains many incorrect/redundant rules due to not only translation variety but also errors in automatic acquisition ((b) in Table 6.1). To overcome this problem, the author proposed the *feedback cleaning* method, which removes incorrect/redundant rules by using the automatic evaluation of machine translation quality. This method regards the removal of the rules as the combinatorial optimization problem of the rules. Specifically, automatic evaluation is used in the combinatorial optimization, and the method searches for the optimal combination as a way to maximize the evaluation scores. BLEU was used for the automatic evaluation. The hill-climbing algorithm, which involves features of this task, was applied to the process of searching for the optimal combination of rules. However, this method requires an evaluation corpus that is different from the training corpus. To avoid this problem, the *N-fold cross-cleaning* method, which uses only the training corpus for cleaning, was proposed. Using the cross-cleaning, MT quality improved by about 11.2% compared with the performance of previous methods.

To summarize in this thesis, an automatic knowledge acquisition method from bilingual corpora was proposed, problems with corpus-based machine translation were identified, and solutions to the problems caused by bilingual corpora and automatic acquisition were advanced proposed in this thesis.

6.2 Remaining Subjects and Future Directions

6.2.1 Word Alignment

The details of word alignment were not discussed in this thesis. However, word alignment is one of the most important techniques in every type of machine translation. For example, hierarchical phrase alignment extracts equivalent phrases from the results of word alignment. Word-level statistical machine translation computes probabilities of the translation model based on the results of word alignment. In addition, even in hand-coded machine translation, word alignment is useful for constructing translation dictionaries.

Quite a few statistical word alignment methods have been proposed (Gale and Church, 1991; Brown et al., 1993; Melamed, 2000; Sumita, 2000; Mihalcea and Pedersen, 2003). These results should be integrated with our methods to further develop automatic construction of knowledge.

6.2.2 Parsing

In this thesis, the parsers contained grammar that was manually developed. This assumed that there is no single best parser in a language because tools or corpora depend on the language activities. However, in the languages that are actively researched, such as English and Japanese, good parsers are available. For example, Charniak (2000) published a parser learned by the maximum entropy in English. In Japanese, Kudo and Matsumoto (2002) published a parser learned by support vector machines (SVMs). We should consider the usage of these parsers. However, the parser of Charniak (2000) output the phrase structure, and the parser of Kudo and Matsumoto (2002) output the dependency structure. These structures have to be unified in order to compute structural similarity.

6.2.3 Bilingual Corpora

Corpus-based machine translation assumes the existence of bilingual corpora. Therefore, we should continue developing corpora. However, as we discussed in Chapter 4, quality of bilingual corpora should be also discussed. The important point is that bilingual sentences are not always symmetric. Assume that the English-to-Japanese and Japanese-to-English MT systems constructed from one bilingual corpus exist. If we translate English sentences into Japanese and translate the Japanese results into English again, the most final results are not equal to the source sentences. This is caused by asymmetry of bilingual sentences. Increasing symmetry is a remaining subject for corpus-based machine translation. Paraphrasing might be an effective way to increase the symmetry (Watanabe et al., 2002).

6.2.4 Machine Translation Engine

Statistical machine translation, in which most systems have been based on word-level translation, is evolving into phrasal translation. For example, Koehn et al. (2003) and Watanabe et al. (2003) proposed methods that translate complex words based on 'phrases' or 'chunks.' Note that 'phrase' denotes a minimal unit of a syntactic phrase and does not involve hierarchy. Yamada and Knight (2001) and Charniak et al. (2003) realized methods that transfer a source word sequence into a target syntactic structure by training with the parsing tree of the target language. Zens and Ney (2003) proposed a word re-ordering system, which constructs tree structure from the viewpoint of whether the word order is inverse or monotone.

Furthermore, some statistical MT systems have absorbed the concept of example-based MT. For example, Och et al. (1999) introduced 'alignment templates,' which are constraints of word-class sequences, to statistical MT. These alignment templates can be regarded as translation rules for example-based MTs. By using these constraints, they achieved idiomatic translation for word-level statistical MT. Marcu (2001) and Watanabe and Sumita (2003) obtained good translations by retrieving similar phrases or sentences to the input sentence from bilingual corpora and modifying the translation based on models of statistical MT.

Example-based MT and statistical MT are indeed not opposite concepts. The key is to adopt merits that each approach offers. Future work will involve constructing an example-based MT engine that exploits the advantages of statistical MT but remains grounded in the syntactic transfer method.

Bibliography

Yasuhiro Akiba, Kenji Imamura, and Eiichiro Sumita. 2001. Using multiple edit distances to automatically rank machine translation output. In *Proceedings of Machine Translation Summit VIII*, pages 15–20.

Eiji Aramaki, Sadao Kurohashi, Satoshi Sato, and Hideo Watanabe. 2001. Finding translation correspondences from parallel parsed corpus for example-based translation. In *Proceedings of Machine Translation Summit VIII*, pages 27–32.

Eiji Aramaki, Sadao Kurohashi, Hideki Kashioka, and Hideki Tanaka. 2003. Word selection for EBMT based on monolingual similarity and translation confidence. In Rada Mihalcea and Ted Pedersen, editors, *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 57–64, Edmonton, Alberta, Canada, May 31. Association for Computational Linguistics.

Marko Auerswald. 2000. Example-based machine translation with templates. In Wolfgang Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, pages 418–427. Springer.

Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. 1991. Aligning sentences in parallel corpora. In *Proceedings of 29th Annual Meeting of the Association for Computational Linguistics (ACL-91)*, pages 169–176.

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

Eugene Charniak, Kevin Knight, and Kenji Yamada. 2003. Syntax-based language models for statistical machine translation. In *Proceedings of Machine Translation Summit IX*, pages 40–46.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2000).*

Michael Collins. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meetings of the Association for Computational Linguistics (ACL-EACL '97)*, pages 16–23.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the HLT Conference*, San Diego, California.

Osamu Furuse and Hitoshi Iida. 1994. Constituent boundary parsing for example-based machine translation. In *Proceedings of COLING-94*, pages 105–111.

Osamu Furuse, Yasuhiro Sobashima, Toshiyuki Takezawa, and Noriyoshi Uratani. 1994. Bilingual corpus for speech translation. In *Proceedings of the AAAI'94 Workshop 'Integration of Natural Language and Speech Processing'*, pages 84–91.

William A. Gale and Kenneth W. Church. 1991. Identifying word correspondences in parallel texts. In *Proceedings of the 4th DARPA Workshop on Speech and Natural Language, Asilomar, CA*, pages 152–157.

Willem-Olaf Huijsen. 1998. Controlled language — an introduction. In *Proceedings of the Second International Workshop on Controlled Language Applications (CLAW-98)*, pages 1–15.

Satoru Ikehara, Satoshi Shirai, Akio Yokoo, and Hiromi Nakaiwa. 1991. Toward an MT system without pre-editing – effects of new methods in ALT-J/E –. In *Proceedings of MT Summit III*, pages 101–106.

Kenji Imamura, Eiichiro Sumita, and Yuji Matsumoto. 2003a. Automatic construction of machine translation knowledge using translation literalness. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, pages 155–162.

Kenji Imamura, Eiichiro Sumita, and Yuji Matsumoto. 2003b. Feedback cleaning of machine translation rules using automatic evaluation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, pages 447–454.

Kenji Imamura. 2001. Hierarchical phrase alignment harmonized with parsing. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS 2001)*, pages 377–384.

Kenji Imamura. 2002. Application of translation knowledge acquired by hierarchical phrase alignment for pattern-based MT. In *Proceedings of the 9th Conference on Theoretical and Methodological Issues in Machine Translation (TMI-2002)*, pages 74–84.

Hiroyuki Kaji, Yuuko Kida, and Yasutsugu Morimoto. 1992. Learning translation templates from bilingual text. In *Proceedings of COLING-92*, pages 672–678.

Genichiro Kikui, Eiichiro Sumita, Toshiyuki Takezawa, and Seiichi Yamamoto. 2003. Creating corpora for speech-to-speech translation. In *Proceedings of EuroSpeech 2003*, pages 381–384.

Mihoko Kitamura and Yuji Matsumoto. 1995. A machine translation system based on translation rules acquired from parallel corpora. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 27–36.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In Marti Hearst and Mari Ostendorf, editors, *HLT-NAACL 2003: Main Proceedings*, pages 127–133, Edmonton, Alberta, Canada, May 27 - June 1. Association for Computational Linguistics.

Taku Kudo and Yuji Matsumoto. 2002. Japanese dependency analyisis using cascaded chunking. In *Proceedings of the 6th Conference on Natural Language Learning 2002 (CoNLL- 2002)*, Taipei.

Daniel Marcu. 2001. Towards a unified approach to memory- and statistical-based machine translation. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pages 386–393.

Yuji Matsumoto, Hiroyuki Ishimoto, and Takehito Utsuro. 1993. Structural matching of parallel texts. In *Proceedings of the 31st Annual Meeting of the ACL*, pages 23–30.

I. Dan Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249, June.

Arul Menezes and Stephen D. Richardson. 2001. A best first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In

Proceedings of the 'Workshop on Example-Based Machine Translation' in MT Summit VIII, pages 35–42.

Adam Meyers, Roman Yangarber, and Ralph Grishman. 1996. Alignment of shared forests for bilingual corpora. In *Proceedings of COLING-96*, pages 460–465.

Adam Meyers, Michiko Kosaka, and Ralph Grishman. 2000. Chart-based translation rule application in machine translation. In *Proceedings of COLING-2000*, pages 537–543.

Rada Mihalcea and Ted Pedersen, editors. 2003. *Building and Using Parallel Texts: Data Driven Machine Translation and Beyond.* HLT-NAACL 2003 Workshop.

Teruko Mitamura and Eric H. Nyberg. 1995. Controlled English for knowledge-based MT: Experience with the KANT system. In *Proceedings of TMI-95.*

Teruko Mitamura and Eric Nyberg. 2001. Automatic rewriting for controlled language translation. In *Workshop on Automatic Paraphrasing: Theories and Applications, NLPRS2001 Post-Conference Workshop*, pages 1–12.

Teruko Mitamura, Eric H. Nyberg, and Jamie G. Carbonell. 1991. An efficient interlingua translation system for multi-lingual document production. In *Proceedings of Machine Translation Summit III*, pages 55–61, Washington, DC.

Makoto Nagao. 1984. A framework of mechanical translation between Japanese and English by analogy principle. In *Artificial and Human Intelligence*, pages 173–180, Amsterdam: North-Holland.

Masaaki Nagata. 1994. A stochastic Japanese morphological analyzer using a forward-DP backward-A* N-best search algorithm. In *Proceedings of COLING-94*, pages 201–207.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 295–302.

Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *Proceedings of the Joint Conference of Empirical Methods in Natural Language Processing and Very*

Large Corpora, pages 20–28, University of Maryland, College Park, MD, June.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In Erhard Hinrichs and Dan Roth, editors, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167.

Susumu Ohno and Masato Hamanishi. 1984. *Ruigo-Shin-Jiten*. Kadokawa, Tokyo. in Japanese.

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

Michael Paul, Kenji Imamura, Eiichiro Sumita, and Seiichi Yamamoto. 2003. Topic-adaptation of pattern-based MT systems using feedback cleaning. In *Proceedings of Recent Advances in Natural Language Processing (RANLP-2003)*, pages 364–368, Borovets, Bulgaria.

Stephen D. Richardson, William B. Dolan, Arul Menezes, and Jessie Pinkham. 2001. Achieving commercial-quality translation with example-based methods. In *Proceedings of Machine Translation Summit VIII*, pages 293–298.

Satoshi Sekine and Ralph Grishman. 1995. A corpus-based probabilistic grammar with only two non-terminals. In *the Fourth International Workshop on Parsing Technology, Prague, Czech.*

Satoshi Shirai, Satoru Ikehara, Akio Yokoo, and Yoshifumi Ooyama. 1998. Automatic rewriting method for internal expressions in Japanese to English MT and its effects. In *Proceedings of the 2nd International Workshop on Controlled Language Applications (CLAW-98)*, pages 62–75.

Harold Somers. 1999. Example-based machine translation. *Machine Translation*, 14:113–157.

Keh-Yih Su, Ming-Wen Wu, and Jing-Shin Chang. 1992. A new quantitative quality measure for machine translation systems. In *Proceedings of COLING-92*, pages 433–439.

Eiichiro Sumita and Hitoshi Iida. 1991. Experiments and prospects of example-based machine translation. In *Proceedings of the 29th ACL*, pages 185–192.

Eiichiro Sumita, Setsuo Yamada, Kazuhide Yamamoto, Michael Paul, Hideki Kashioka, Kai Ishikawa, and Satoshi Shirai. 1999. Solutions to problems inherent in spoken-language translation: The ATR-MATRIX approach. In *Machine Translation Summit VII*, pages 229–235.

Eiichiro Sumita. 2000. Word alignment using a matrix. In *Proceedings of PRICAI2000*, page 821. Springer.

Eiichiro Sumita. 2003. An example-based machine translation system using DP-matching between word sequences. In Michael Carl and Andy Way, editors, *Recent Advances in Example-based Machine Translation*, pages 189–209. Kluwer Academic Publishers.

Toshiyuki Takezawa and Tsuyoshi Morimoto. 1997. Dialogue speech recognition using syntactic rules based on subtrees and preterminal bigrams. *Systems and Computers in Japan*, 28(5):22–32.

Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. 2002. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 147–152.

Masao Utiyama and Hitoshi Isahara. 2003. Reliable measures for aligning Japanese-English news articles and sentences. In Erhard Hinrichs and Dan Roth, editors, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 72–79.

Wolfgang Wahlster, editor. 2000. *Verbmobil: Foundations of Speech-to-Speech Translation.* Springer-Verlag Berlin Heidelberg.

Taro Watanabe and Eiichiro Sumita. 2003. Example-based decoding for statistical machine translation. In *Proceedings of Machine Translation Summit IX*, pages 410–417.

Hideo Watanabe, Sadao Kurohashi, and Eiji Aramaki. 2000. Finding structural correspondences from bilingual parsed corpus for corpus-based translation. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, pages 906–912.

Taro Watanabe, Mitsuo Shimohata, and Eiichiro Sumita. 2002. Statistical machine translation on paraphrased corpora. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1954–1957.

Taro Watanabe, Eiichiro Sumita, and Hiroshi G. Okuno. 2003. Chunk-based statistical translation. In Erhard Hinrichs and Dan Roth, editors, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 303–310.

Dekai Wu. 1995. An algorithm for simultanteously bracketing parallel texts by aligning words. In *Proceedings of the 33rd Annual Meeting of the Assoc. for Computational Linguistics (ACL-95)*, pages 244–251.

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

Setsuo Yamada, Kazuhide Yamamoto, and Hitoshi Iida. 1998. Word selection on cooperative integrated machine translation. In *Proceedings of The 4th Annual Meeting of The Association for Natural Language Processing*, pages 508–511. in Japanese.

Kaoru Yamamoto and Yuji Matsumoto. 2000. Acquisition of phrase-level bilingual correspondence using dependency structure. In *Proceedings of COLING-2000*, pages 933–939.

Kazuhide Yamamoto, Jun Kawai, Eiichiro Sumita, and Osamu Furuse. 1997. Morphological analysis utilizing n-gram of mixed category. In *Proceedings of the 54th Annual Meeting of Information Processing Society of Japan, 1C-02*. in Japanese.

Keiji Yasuda, Fumiaki Sugaya, Toshiyuki Takezawa, Seiichi Yamamoto, and Masuzo Yanagida. 2001. An automatic evaluation method of translation quality using translation answer candidates queried from a parallel corpus. In *Proceedings of Machine Translation Summit VIII*, pages 373–378.

Keiji Yasuda, Fumiaki Sugaya, Toshiyuki Takezawa, Seiichi Yamamoto, and Masuzo Yanagida. 2003. Applications of automatic evaluation methods to measuring a capability of speech translation system. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, pages 371–378.

Richard Zens and Hermann Ney. 2003. A comparative study on reordering constraints in statistical machine translation. In Erhard Hinrichs and Dan Roth, editors, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 144–151.

List of Publications

Journals

(1) Kenji Imamura, Eiichiro Sumita, and Yuji Matsumoto. 2004b. Feedback cleaning of machine translation rules using automatic evaluation. In *IPSJ (Information Processing Society of Japan) Journal*, Vol. 45, Number 8. in Japanese. to appear.

(2) Kenji Imamura, Eiichiro Sumita and Yuji Matsumoto. 2004a. Automatic construction of machine translation knowledge using translation literalness. In *Journal of Natural Language Processing*, Vol. 11, Number 2, pages 85–99. in Japanese.

(3) Kenji Imamura. 2002. Hierarchical phrase alignment harmonized parsing. In *Journal of Natural Language Processing*, Vol. 9, Number 5, pages 23–42. in Japanese.

(4) Kenzi Imamura, Minoru Ohyama, and Yukio Hashida. 1987. A sentence generation method of a route guidance system using a text-to-speech synthesizer. In *the Transactions of the Institute of Electronics, Information and Communication Engineers*, Vol. J71-D, No.5, pp.926-929. Short Note. in Japanese.

Journal (co-author)

(5) Taro Watanabe, Kenji Imamura, Eiichiro Sumita, and Hiroshi G. Okuno. 2004. Statistical machine translation using hierarchical phrase alignment. *IEICE Transactions on Information and Systems, Pt. 2*, Vol. J87-D-II, No. 4, pp. 978–986. in Japanese.

International Conferences & Workshops

(6) Kenji Imamura, Hideo Okuma, Taro Watanabe, and Eiichiro Sumita. 2004. Example-based machine translation based on syntactic transfer with statistical models. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland, to appear.

(7) Kenji Imamura, Eiichiro Sumita, and Yuji Matsumoto. 2003b. Feedback cleaning of machine translation rules using automatic evaluation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, Sapporo, Japan, pages 447–454.

(8) Kenji Imamura, Yasuhiro Akiba, and Eiichiro Sumita. 2003. Automatic expansion of equivalent sentence set based on syntactic substitution. In *Conference Conpanion Volume of Human Language Technology Conference 2003 (HLT-NAACL 2003)*, Edmonton, Canada, pages 37–39.

(9) Kenji Imamura, Eiichiro Sumita, and Yuji Matsumoto. 2003a. Automatic construction of machine translation knowledge using translation literalness. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, Budapest, Hungary, pages 155–162.

(10) Kenji Imamura and Eiichiro Sumita. 2002. Bilingual corpus cleaning focusing on translation literality. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, Denver, U.S.A., pages 1713–1716.

(11) Kenji Imamura. 2002. Application of translation knowledge acquired by hierarchical phrase alignment for pattern-based MT. In *Proceedings of the 9th Conference on Theoretical and Methodological Issues in Machine Translation (TMI-2002)*, Keihanna, Japan, pages 74–84.

(12) Kenji Imamura. 2001b. Hierarchical phrase alignment harmonized with parsing. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS-2001)*, Tokyo, Japan, pages 377–384.

(13) Kenji Imamura. 2001a. A Hierarchical Phrase Alignment from English and Japanese Bilingual Text. In *Proceedings of 2nd International Conference, CICLing 2001*, Mexico City, Mexico, pages 206–207.

(14) Kenzi Imamura and Yoshifumi Ooyama. 1995. Identifying user model of the sender and receiver in a message domain. In *Proceedings of PACLING-II*, Brisbane, Australia.

(15) Kenzi Imamura. 1993. Rules of rhythm construction in Tanka and judgment method of rhythm. In *Proceedings of the 2nd Natural Language Processing Pacific Rim Symposium (NLPRS-93)*, Kitakyuushuu, Japan, pages 406–410.

International Conferences & Workshops (co-author)

(16) Michael Paul, Kenji Imamura, Eiichiro Sumita, and Seiichi Yamamoto. 2003. Topic-adaptation of pattern-based MT systems using feedback cleaning. *Proceedings of the International Conference: Recent Advances in Natural Language Processing (RANLP-2003)*, pages 364–368.

(17) Eiichiro Sumita, Yasuhiro Akiba, Takao Doi, Andrew Finch, Kenji Imamura, Michael Paul, Mitsuo Shimohata, and Taro Watanabe. 2003. A corpus-centered approach to spoken language translation. In *EACL 2003 Conference Companion*, Budapest, Hungary, pages 171–174.

(18) Eiichiro Sumita, Yasuhiro Akiba, and Kenji Imamura. 2002. Reliability measures for translation quality. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, Denver, U.S.A., pages 1893–1896.

(19) Kazutaka Takao, Kenji Imamura, and Hideki Kashioka. 2002. Comparing and extracting paraphrasing words with 2-way bilingual directionaries. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, Canaria, Spain, pages 1016-1022.

(20) Taro Watanabe, Kenji Imamura, and Eiichiro Sumita. 2002. Statistical MT based on hierarchical phrase alignment. In *Proceedings of the 9th Conference on Theoretical and Methodological Issues in Machine Translation (TMI-2002)*, Keihanna, Japan, pages 188–198.

(21) Setsuo Yamada, Kenji Imamura, and Kazuhide Yamamoto. 2002. Corpus-assisted expansion of manual MT knowledge. In *Proceedings of the 9th Conference on Theoretical and Methodological Issues in Machine Translation (TMI-2002)*, Keihanna, Japan, pages 199–208.

(22) Yasuhiro Akiba, Kenji Imamura, and Eiichiro Sumita. 2001. Using multiple edit distances to automatically rank machine translation output. In

Proceedings of Machine Translation Summit VIII, Santiago, Spain, pages 15–20.

(23) Yukiko I. Nakano, Kenji Imamura, and Hisashi Ohara. 2000. Taking account of the user's view in 3D multimodal instruction dialogue. In *Proceedings of COLING 2000*, Saarbrucken, Germany, pages 572-578.

(24) Yoshifumi Ooyama, Tsuneaki Kato, and Kenzi Imamura. 1995. Japanese kanji name input system using spoken-style transcibed Japanese analysis. In *Proceedings of the 11th Conference on Artificial Intelligence for Applications.*

Local Conferences & Workshops

(25) Kenji Imamura, Eiichiro Sumita, and Yuji Matsumoto. 2004. 機械翻訳自動評価指標の比較 "Comparison study among automatic evaluation metrics of machine translation quality." In *Proceedings of 10th Annual Meeting of Natural Language Processing (NLP 2004)*, pp. 452-455, in Japanese.

(26) Kenji Imamura and Eiichiro Sumita. 2002. 直訳性に着目した対訳コーパスフィルタリング "Bilingual corpus filtering based on translation literality." In *Proceedings of the 1st Forum on Information Technology (FIT 2002)*, Vol. E, pp. 185-186. in Japanese.

(27) Kenji Imamura, Yasuhiro Akiba, and Eiichiro Sumita. 2001. 階層的句アライメントを用いた日本語翻訳文の換言 "Paraphrasing of Japanese translation set using hierarchical phrase alignment." In *Proceedings of Workshop Program of the 7th Annual Meeting of the Association for Natural Language Processing*, pp. 15-20. in Japanese.

(28) Kenji Imamura and Yoshifumi Ooyama. 1994. メッセージにおける送り手・受け手のユーザモデル抽出 "Identifying user model of the sender and receiver in a message domain." *IPSJ SIG Notes*, 94-NL-103, pp.73-80. in Japanese.

(29) Kenji Imamura, Motoyuki Horii, and Yoshifumi Ooyama. 1993. 言語表現を利用したメッセージの送り手の性別判定 "A sex identification method of a message sender using linguistic characteristics." In *Proceedings of the 46th IPSJ Annual Meetings*, 3B-03. in Japanese.

(30) Kenji Imamura, Ken'ichi Kuroda, and Yoshifumi Ooyama. 1992. メッセージの女性→男性表現変換の検討 "A message translation method from female

to male expression.” In *Proceedings of the 44th IPSJ Annual Meetings*, 3Q-7. in Japanese.

(31) Kenji Imamura and Yoshifumi Ooyama. 1991. メッセージ検索方式の検討 “Research of message database retrieval.” In *1991 Spring National Convention Record, IEICE*, 6-109. in Japanese.

(32) Kenji Imamura, Hideyuki Tsuchiya, and Yoshifumi Ooyama. 1989. DB検索におけるメニュー入力と自然語入力時間の比較 “A comparison of the efficiency between menu selection and natural language input method in database retrieval.” In *Proceedings of the 39th IPSJ Annual Meetings*, pp.1451-1452. in Japanese.

(33) Kenji Imamura, Minoru Ohyama, and Yukio Hashida. 1987b. 音声による道案内システム〜自然な案内文の合成 “A course guidance system using a speech synthesizer – synthesis of natural guidance.” In *National Conference Record, 1987: Information and Systems, IEICE*, 1-145. in Japanese.

(34) Kenji Imamura, Minoru Ohyama, and Yukio Hashida. 1987a. 音声による道案内システム “A course guidance system using a text to voice converter.” In *National Conference Record, 1987, IEICE*, 6-621. in Japanese.

Local Conferences & Workshops (co-author)

(35) Kazutaka Takao, Kenji Imamura, and Hideki Kashioka. 2002. 英和／和英辞典を用いた英語換言語の抽出 “Extracting Paraphrasing English Words with E-J and J-E Dictionaries.” In *Proceedings of the 8th Annual Meeting of the Association for Natural Language Processing*, pp. 128-131. in Japanese.

(36) Yasuhiro Akiba, Kenji Imamura, and Eiichiro Sumita. 2001. 複数編集距離を用いた機械翻訳の自動評価 “Automatic evaluation of translation systems by using multiple edit distances.” In *Proceedings of the 63rd IPSJ Annual Meetings*, pp.2-257-258. in Japanese.

(37) Kazutaka Takao, Mitsuo Shimohata, Kenji Imamura, and Hideki Kashioka. 2001. 和英／英和辞典のカバレッジの比較と応用 “Comparison of coverage and application between Japanese-to-English and English-to-Japanese dictionaries.” In *Proceedings of the 7th Annual Meeting of the Association for Natural Language Processing*, pp. 58-61. in Japanese.

(38) Yukiko I. Nakano and Kenji Imamura. 1999b. 協調的対話機構と3次元仮想学習環境の統合 “Integrating collaborative dialogue system with 3D virtual

learning environment.” In *IEICE Technical Report*, ET99-17, pp.57-64. in Japanese.

(39) Yukiko I. Nakano and Kenji Imamura. 1999a. 3次元仮想環境における教示対話生成システム “Instruction dialogue system in 3D virtual environment.” In *Proceedings of the 13rd Annual Conference of JSAI.* in Japanese.

(40) Rintaro Sunaba, Motoyuki Horii, Kenji Imamura, and Yoshifumi Ooyama. 1993. メッセージ種別判定方式の検討 “A method of message classification.” In *1993 Spring National Convention Record, IEICE.* in Japanese.

(41) Motoyuki Horii, Kenji Imamura, and Yoshifumi Ooyama. 1992. 文のおもしろさを決定する言語的な要因の分析 “Determining the humorousness of Japanese telegram sentences.” In *Proceedings of the 45th IPSJ Annual Meetings*, 6G-07. in Japanese.

(42) Yoshifumi Ooyama and Kenji Imamura. 1991. 住所入力支援方式の検討 “A Japanese address input method.” In *1991 Spring National Convention Record, IEICE*, 6-106. in Japanese.

(43) Motoyuki Horii, Kenji Imamura, Tsuneaki Kato, and Yoshifumi Ooyama. 1990. メッセージにおける言語表現の分析とその生成 “A method for generating ‘messages’ using the analysis of telegrams.” *IPSJ SIG Notes*, NL78-14, pp.105-112. in Japanese.

Patents

(44) Kenji Imamura and Yoshifumi Ooyama. 1993. 自然語文解析装置、文リズムパターン選択装置及び文生成装置 “Natural language analysis system, sentence rhythm pattern selection system, and sentence generation system.” *Japanese Patent No. 3341176.*

(45) Kenji Imamura, Motoyuki Horii, and Yoshifumi Ooyama. 1992. メッセージの発信者・受信者情報判定装置 “Identification of sender and receiver information in messages.” *Japanese Patent No. 3131934.*

(46) Kenji Imamura, Yoshifumi Ooyama, Motoyuki Horii, Tsuneaki Kato, and Masahiro Oku. 1991. メッセージ検索装置 “Message retrieval system.” *Japanese Patent No. 3090285.*

(47) Kenji Imamura, Yoshifumi Ooyama, Tsuneaki Kato, and Masahiro Oku. 1990. 言語表現の特徴判定装置 “Identification system of language expression characteristics.” *Japanese Patent No. 3043038.*

(48) Kenji Imamura and Yoshifumi Ooyama. 1990. データベース検索解表示装置 “Viewing system of database retrieval results.” *Japanese Patent No. 02749420.*

Patents (co-author)

(49) Yoshifumi Ooyama, Kenji Imamura, and Masanobu Higashida. 1992. 予約案内装置 “Reservation guidance system.” *Japanese Patent No. 2945928.*

(50) Tsuneaki Kato, Yoshifumi Ooyama, and Kenji Imamura. 1990. かな漢字変換装置 “Kana-Kanji conversion system.” *Japanese Patent No. 2759123.*

(51) Yoshifumi Ooyama, Kenji Imamura, and Tsuneaki Kato. 1990. 辞書作成装置 “Dictionary construction system.” *Japanese Patent No. 02628775.*

(52) Hiroshi Matsuo, Yoshifumi Ooyama, and Kenji Imamura. 1989. 学習機能付き自然文意味解析処理装置 “Semantic analysis system of natural language including learning functions.” *Japanese Patent No. 2759123.*

(53) Minoru Ohyama, Kenji Imamura, and Yukio Hashida. 1988. 音声認識システム “Speech recognition system.” *Japanese Patent No. 2080108.*