NAIST-IS-DD0261014

Doctoral Dissertation

Acquiring Paraphrases from Corpora and Its Application to Machine Translation

Mitsuo Shimohata

September 15, 2004

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

A Doctoral Dissertation
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

Mitsuo Shimohata

Thesis Committee:

Professor Yuji Matsumoto	(Supervisor)
Professor Shunsuke Uemura	(Co-supervisor)
Professor Kiyohiro Shikano	(Co-supervisor)

Acquiring Paraphrases from Corpora and Its Application to Machine Translation*

Mitsuo Shimohata

Abstract

A natural language contains various paraphrases, that is, superficially different expressions that share the same meaning. Such a wide variety of paraphrases reflects the rich expressiveness of natural language, while causing difficulty in natural language processing applications, such as machine translation (MT). For MT, this variety reduces the coverage of translatable input sentences and complicates language too much to comprehend every possible variation. Unfortunately, existing resources for paraphrases do not adequately deal with the difficulty because their paraphrase knowledge only covers general areas and has little effect on uses for specific domains and applications.

This thesis describes corpus-based paraphrase acquisition and its application to MT. We propose two paraphrase acquisition methods: lexical paraphrases and sentential paraphrases, each of which has its own advantages. Both methods are based on shallow analysis, and rely on a corpus but no other resource. The achievements described in this thesis consist of three parts: analysis of manual paraphrases, automatic acquisition of lexical paraphrases, and similar sentence retrieval, which corresponds to sentential paraphrasing.

First, we describe two analyses of human paraphrases to clarify the following questions: (1) what types of paraphrases are dominant? and (2) how can human paraphrases be effective for MT? These investigations suggest that lexical paraphrasing and sentential paraphrasing are dominant in travel conversation domains.

Second, we describe a method for extracting lexical paraphrases from a parallel corpus. This method has two advantages: (1) it acquires not only synonymous content

*Doctoral Dissertation, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD0261014, September 15, 2004.

words but also function word related synonymous expressions, and (2) it extracts paraphrases from a translingual viewpoint. Extracted paraphrases simplify texts in a given corpus by unifying paraphrases into a single expression. This unification of paraphrases improves the performance of corpus-based MT. We demonstrate the effect of paraphrase unification on two different corpus-based MT systems: example-based MT and statistical MT.

Finally, we describe a method for retrieving a similar sentence from a monolingual corpus. This method provides sentential paraphrases from a monolingual corpus, and it has two advantages: (1) it relies on a monolingual corpus which is easy to prepare, and (2) it acquires paraphrases whose differences go beyond the lexical level. We selected metrics based on N-gram co-occurrence to measure sentence similarity after conducting a comparative study among three major metrics. This similar sentence retrieval technique is applied to MT in two ways. One is a pre-edit of an input sentence. If a given input sentence is found to be difficult to translate, a similar translatable sentence is retrieved and given to the MT instead of the original input sentence. The other is an example-based rough translation system. We used our similarity measure method with an EBMT system. Our retrieval method gives an EBMT system a wider coverage of translatable sentences because the retrieval is robust. We named the translation method Rough Translation.

Keywords:

Paraphrase, Machine Translation, Parallel Corpus, Speech Translation, Pre-edit, N-gram Co-occurrence

Acknowledgments

When I think back on my research career, I am grateful for having the great fortune of having been able to carry out my research activities to this day. This work has been supported by many people, with my effort being but a small part. Here, I would like to express my gratitude for those who have supported me.

I am sincerely grateful to Professor Yuji Matsumoto, the advisor for my doctoral thesis. Professor Matsumoto gave me many useful comments from his wide range of knowledge. I am grateful to the members of his laboratory, who share the professor's diligence. Their earnest attitude toward NLP research stimulated me greatly, and my discussions with them were useful.

I am also grateful to the dissertation committee, including Professor Shunsuke Uemura and Professor Kiyohiro Shikano. They gave me highly useful comments that allowed me to consider the fundamental positioning of my research.

The achievements described in this thesis were mainly supported by the Advanced Telecommunication Research Laboratories (ATR). Dr. Seiichi Yamamoto, director of ATR-SLT, gave me the chance to conduct research in automatic paraphrasing. Mr. Satoshi Shirai and Dr. Hiromi Nakaiwa, former heads of Department 3, supervised my research. Dr. Eiichiro Sumita guided me in my paraphrasing research and encouraged me constantly. The members of the NLP Department, Mr. Yasuhiro Akiba, Mr. Takao Doi, Dr. Andrew Finch, Dr. Kenji Imamura, Dr. Hideki Kashioka, Dr. Kiyonori Ohtake, Mr. Michael Paul, Dr. Taro Watanabe, Dr. Kazuhide Yamamoto, and Mr. Nobutaka Yoshioka, all supported my research efforts.

I would like to thank the staff of Oki Electric Industry. Yoshitaka Hirogaki, Makoto Morito, and Toshihisa Nakai supported me as I carried out my activities at ATR. Junichi Fukumoto, Atsushi Ikeno, Mihoko Kitamura, Fumito Masui, Toshihiko Miyazaki, Toshiki Murata, and Hideki Yamamoto imparted me a wide range of knowledge and

invaluable help in my NLP research.

I especially wish to thank my family. My parents, Tatsumi Shimohata and Sonoe Shimohata, helped to develop my personality and supported my academic education. Sayori Shimohata, my wife, a friendly advisor in my research and the caring mother of our three children, has been contributing most of her energy for childcare and sacrificing her own research activities. Her contribution has encouraged and supported me greatly. My lovely children, Inori, Madoka, and Nodoka, also encouraged me. Their cute smiles allowed me to relax, while they kept me busy.

Finally, I would be very pleased if this thesis repaied even a small part of the many contributions made by my supporters.

Contents

List of Figures **x**

List of Tables **xi**

1 Introduction **1**

1. Paraphrase 1
2. Research on Automatic Paraphrase Acquisition 3
3. Application of Paraphrasing 4
4. Research Objectives 4
5. Thesis Outline 5

2 Research Resources **9**

1. Corpora 9
   - 1.1 Transcribed Dialog Corpus 10
   - 1.2 Basic Travel Expression Corpus 11
   - 1.3 Difference of the Two Corpora 11
2. Corpus-based Machine Translation Systems 13
   - 2.1 Example-based Machine Translation System 13
   - 2.2 Statistical Machine Translation System 14

3 Analysis of Manual Paraphrasing **15**

1. Paraphrasing Types 15
   - 1.1 Sentential Paraphrase 16
   - 1.2 Phrasal Paraphrase 17
   - 1.3 Lexical Paraphrase 17

2. Construction of Paraphrase Corpus 18
3. Analysis of Paraphrased Sentences 20
4. Related Work 21
5. Conclusion 22

4 Building a Paraphrase Corpus for Speech Translation 23
1. Basic Idea of Paraphrasing 23
1.1 Sentence Length Metric 24
1.2 Threshold for Long Sentences 25
2. Paraphrasing Methods 25
2.1 Segment Paraphrasing 25
2.2 Summary Paraphrasing 26
2.3 Plain Paraphrasing 26
3. Experiment 27
3.1 Experimental MT Systems 27
3.2 Translation Quality 28
3.3 Cross Perplexity 29
3.4 Positive/Negative Paraphrasing between Original and Paraphrased Sentences 30
4. Conclusions 31

5 Extracting Lexical Paraphrases from a Parallel Corpus 33
1. Features of Synonymous Expressions 34
1.1 Synonymous from the Translingual Viewpoint 34
1.2 Expressions Including Context 35
1.3 Function Word Related Expressions 36
1.4 Lexical Synonymous Expressions 36
2. Extraction of Synonymous Expressions 37
2.1 Basic Extraction Process 37
2.2 Iteration 40
3. Experiment 41
3.1 Corpus 41
3.2 Extracted Synonymous Expressions 42
3.3 Application to EBMT 45

3.4 Application to SMT 47
4. Related Work 48
4.1 Thesauri 48
4.2 Automatic Acquisition of Lexical Synonyms 49
5. Comparative Experiments on Various Language Pairs 49
5.1 Evaluation Method 50
5.2 Results 51
6. Conclusion 52

6 Retrieving a Similar Sentence from a Monolingual Corpus 53
1. Method for Measuring Similarity between Two Sentences 55
1.1 Overview of Automatic Evaluation of Machine Translation . . . 55
1.2 Basic Methods 56
1.3 Additional Conditions 57
1.4 Comparing Precision of Each Method and Additional Conditions 59
2. Filtering Retrieved Sentences 60
2.1 Number of Missing Content Words 61
2.2 Number of Common Content Words 61
2.3 Results with Filtering 62
3. Experiment on Application to Machine Translation 63
3.1 Experimental MT System 63
3.2 Results 64
4. Conclusions 65

7 Rough Translation based on Similar Sentence Retrieval 67
1. Related Work 68
2. Difficulties of Applying EBMT to S2ST 68
2.1 Translation Degradation by Input Length 68
3. Retrieving Meaning-equivalent Sentences for Rough Translation 69
3.1 Meaning-equivalent Sentences 70
3.2 Basic Idea of Retrieval of Meaning-equivalent Sentences 71
3.3 Features for Retrieval 71
3.4 Retrieval and Ranking 73
4. Modification 74

5. Experiment 75
5.1 Test Data 75
5.2 Compared Methods for Meaning-equivalent Sentence Retrieval . 77
5.3 Accuracy of Meaning-equivalent Sentence Retrieval 77
5.4 Translation Accuracy 78
6. Conclusion 79

8 Conclusion **81**
1. Summary 81
2. Future Work 83

Bibliography **85**

List of Publications **93**

List of Figures

1.1 Examples of Paraphrases 1
1.2 Thesis Overview 6

2.1 Sample Sentences from TDC 10
2.2 Sample Sentences from BTEC 12
2.3 Structure of EBMT System 13

3.1 Examples of Sentential Paraphrasing 16
3.2 Examples of Phrasal Paraphrasing 17
3.3 Examples of Lexical Paraphrasing 18
3.4 Examples of Paraphrased Sentences 19
3.5 Overlap of Paraphrased Sentences between Paraphrasers 21

4.1 Ratio of Content Words in BTEC Corpus 25
4.2 Examples of Paraphrasing 26
4.3 Overview of Experiment 27

5.1 Synonyms from the Translingual Viewpoint 35
5.2 Examples of Lexical Difference 37
5.3 Synonymous Sentence Group 38
5.4 Extraction of Synonymous Expression Pairs 39
5.5 Integrating SS Group by Unification 41
5.6 Extracted Synonymous Expressions by Iteration 43
5.7 Example of English Extracted Synonymous Expression Groups 44
5.8 Example of Japanese Extracted Synonymous Expression Groups 44
5.9 POS Distribution of Extracted Synonymous Expressions 45

5.10 Experimental EBMT System 46
5.11 Examples of Paraphrased Sentence Evaluation 50
5.12 Results for Various Language Pairs 51

6.1 Improving Translation by Similar Sentence Retrieval 54
6.2 Examples of Similarity Evaluation 59
6.3 Precision by Number of Missing Content Words 61
6.4 Accuracy by Number of Common Content Words 62
6.5 Retrieval Precision with Filtering 63
6.6 Overview of Experiment 64

7.1 Distribution of Untranslated Input Sentences by Length 69
7.2 Examples of Unimportant Information 70
7.3 Sentences and Their Modality and Tense 73
7.4 Example of Conditions for Ranking 75
7.5 Replacement of Synonymous Words 76
7.6 Retrieval Accuracy 77
7.7 Translation Accuracy 78

List of Tables

2.1 Statistics of SLDB 11
2.2 Statistics of BTEC 11
2.3 Cross Perplexity 12

3.1 Number of Acquired Paraphrased Sentences by Each Type 20
3.2 Expansion Ratio for Each Paraphrasing Type 21

4.1 Translation Quality 28
4.2 Cross-perplexity 29
4.3 Positive/Negative Paraphrasing Cases 30

5.1 Corpus for Synonymous Expression Extraction 42
5.2 Extracted Synonymous Expression Groups 42
5.3 Expansion of Translatable Inputs by Unifying Synonymous Expressions 46
5.4 Evaluation of Translation Quality 47
5.5 Difference in Translation Quality on SMT by Unifying Synonymous Expressions 48

6.1 Precision by Basic Methods and Additional Conditions 60
6.2 Translation Quality of Original and Retrieved Sentences 65

7.1 Clues for Discriminating Modalities in Japanese 72
7.2 Statistics of the Corpora 76
7.3 Overall Accuracy with Conversational Data 79

Chapter 1

Introduction

1. Paraphrase

According to the Oxford English Dictionary, paraphrase means "*to express the meaning of something using different words.*" Indeed, we can express a thing in various ways according to the speaker's aim. Figure 1.1 shows examples of paraphrases. Sentences 1-1 and 1-2 only differ in the words "card" and "credit card." Both words have the same meaning in this context. The word "card" is more common than "credit card," although its denotation is so ambiguous that it can suggest different objects such as post cards or playing cards. Sentences 2-1 and 2-2 share the same meaning but differ in voice. Sentence 2-2 emphasizes the object "book" by using it as the subject. Sentences

1-1	Can I pay with this **card**?
1-2	Can I pay with this **credit card**?
2-1	Tom gave Mary a book.
2-2	A book was given to Mary by Tom.
3-1	**Please** reply to this email.
3-2	**Would you** reply to this email?
3-3	**I would appreciate if you would** reply to this email.

Figure 1.1. Examples of Paraphrases

3-1, 3-2, and 3-3 express the same request to the recipient of the letter but differ in level of politeness. A sender can choose the proper expression by giving consideration to his or her relationship with the recipient.

Here, we clarify the definition of "paraphrase" in terms of range and type. Although the meaning of "paraphrase" implies that a given *sentence* is converted to make it *easier to understand*, we do not adhere to this common implication. Our paraphrases involve various ranges: lexical, phrasal, and sentential. In addition, our paraphrases are not confined to making a sentence easier to understand. Therefore, synonyms, which correspond to lexical-level paraphrases, are also our target paraphrases. This generic definition of paraphrase is widely applied in the paraphrase research field.

The paraphrasing phenomenon is intriguing and has been the object of many research efforts. Sociolinguistics is a linguistic research field that investigates language variations by such social factors as ethnicity, social class, sex, geography, and age (Nakao et al., 1997).

In Japan, many studies of paraphrases concern honorifics. (Ide et al., 1986) reported a comparative investigation on honorific usage between Japanese and American people. (Marumoto et al., 2003) and (Shirado et al., 2003) investigated differing impressions of various honorific expressions. (Tanaka, 2004) compared the request forms used by Japanese with those used by Germans.

Other studies on paraphrases have also been reported. The authors investigated the variety of human paraphrasing by focusing on three types (described in Section 3). (Sugaya et al., 2002) and (Kinjo et al., 2003) investigated how Japanese translations can be generated from an English sentence. (Fujita and Inui, 2003) analyzed how lexical and structural paraphrasing patterns cause errors.

From a practical viewpoint, paraphrasing brings variation to a language and thus causes difficulty in natural language processing (NLP) applications, such as machine translation (MT), information retrieval (IR), and summarization. Here, we illustrate the difficulty with a simple example. Suppose that an NLP module knows "luggage" but does not know "baggage." When the module conducts MT, it can translate "My luggage is missing" but cannot translate "My baggage is missing." When the module conducts IR and receives "luggage" as a query word, it ignores the data that contains "baggage" from retrieval, and some useful data are not retrieved.

To overcome this difficulty, knowledge of paraphrases has been constructed as the-

sauri and dictionaries. Currently, many manually constructed resources are available. WordNet (Fellbaum, 1998) and Roget's Thesaurus (Roget, 1946) are well-known English Thesauri, and Goi-Taikei (Ikehara et al., 1997) and Bunrui-Goi-Hyo (The National Institute for Japanese Language, 1964) are well-known Japanese thesauri.

Unfortunately, manually constructed resources are not sufficient for NLP applications (Frakes and Baeza-Yates, 1992) since these resources aim at covering general areas and thus have limited effectiveness in specific domains and applications.

2. Research on Automatic Paraphrase Acquisition

Automatic paraphrase acquisition has been developed to overcome the shortcomings of existing resources. In the IR research field, automatic paraphrase acquisition is known as "term clustering" or "word clustering" (Kowalski, 1997; Frakes and Baeza-Yates, 1992). Term clustering provides more benefit to the recall process than to precision. Basically, the commonality of two terms is determined from co-occurrence in the same documents or sentences. However, a serious problem is that such paraphrases mostly deal with synonyms, rarely considering function word related expressions. Obviously, function word related paraphrases are abundant, as exempilfied in Figure 1.1.

A significant point for acquiring function word related paraphrases is how to obtain synonymous text chunks.[1] Much research has attempted to obtain synonymous sentences by binding translations that share the same source sentence (Shimohata et al., 2003b; Barzilay and McKeown, 2001; Pang et al., 2003; Ohtake and Yamamoto, 2003). This approach can provide various types of paraphrases, such as lexical and syntactic paraphrases. Other research has tried to obtain them by binding sentences that share comparable content words (Shimohata et al., 2004b; Shinyama et al., 2002; Imamura et al., 2001; Jacquemin et al., 1997). This approach mainly acquires syntactic paraphrasing rules. Other research has used supplementary knowledge such as dictionaries (Fujita and Inui, 2001; Kaji et al., 2002a), case frame (Kaji et al., 2002b; Torisawa, 2001), and dependency relations (Yamamoto, 2002).

[1] These correspond to sentences, phrases, and others

3. Application of Paraphrasing

Paraphrases vary language expressions and thus cause difficulty in NLP. Therefore, comprehension of paraphrases could benefit many NLP applications. For MT, paraphrasing of an input sentence to normalize various expressions is known as a "pre-edit" and has been tackled by many researchers (Shimohata et al., 2003b; Shimohata et al., 2004b; Shirai et al., 1995; Kim and Ehara, 1994; Yoshimi et al., 2000; Ohtake and Yamamoto, 2001). For automatic MT evaluation, expansion of reference translation by paraphrasing has been reported (Finch et al., 2004). Comprehension of paraphrases has been tackled in other NLP application such as summarization (Barzilay and McKeown, 2001; Kondo and Okumura, 1997), information extraction (Shinyama et al., 2002), and Q&A (Takahashi et al., 2003; Duclaye et al., 2003). Paraphrases also benefit human language activities by improving text readability (Inui et al., 2003; Carroll et al., 1999).

4. Research Objectives

The aim of this thesis is to acquire paraphrases from a corpus and utilize them to improve MT performance. Surprisingly, corpus-based automatic paraphrasing for MT has not yet been attempted. All previous works on automatic pre-edits require manually constructed paraphrasing rules (Shirai et al., 1995; Kim and Ehara, 1994; Yoshimi et al., 2000; Ohtake and Yamamoto, 2001). In addition, their paraphrasing types are limited to sentence segmentation and deletion of redundant expressions; our methods, on the other hand, consider other types of paraphrasing.

Our extraction method relies only on a tagged corpus, i.e., we do not use other linguistic resources such as dictionaries and parsers. We adopted this approach for the following two reasons: (1) We want to clarify the independent effect of paraphrasing on MT. Combinational use of other resources obscures the paraphrasing effect. (2) This approach is favorable for practical use. Independence from rich resources such as parsers and dictionaries permits both fast processing and easy preparation. As a matter of course, our method can attain a synergy effect in combination with other resources.

We propose two methods for paraphrase extraction: lexical paraphrasing and sentential paraphrasing. The former is based on a parallel corpus. The extracted lexical paraphrases simplify a given corpus by unifying synonymous expressions into a single

expression. The method can also extract function word related synonymous expressions and includes expressions that are synonymous from a translingual viewpoint.

Similar sentence retrieval, which corresponds to a sentential paraphrase, uses a monolingual corpus. If an MT system cannot translate an input sentence, the most similar sentence is retrieved from a monolingual corpus and given to the MT system to obtain a translation. We set as a metrics of similarity measurement an N-gram overlap from the results of comparative experiments.

It must be noted that we do not deal with paraphrases of metaphors and idioms. These paraphrases share few words with the original sentence. Since our paraphrase extraction relies on superficial similarity between sentences, these types of paraphrases are beyond the scope of this work.

We have investigated manual paraphrasing in travel conversation, which is our target domain. This investigation clarifies the characteristic of paraphrasing and suggests guidelines for approaching a variety of paraphrases for MT.

5. Thesis Outline

The achievements described in this thesis involve three technical topics: analysis of human paraphrasing, automatic extraction of lexical paraphrases, and similar sentence retrieval (i.e., sentential paraphrasing). These achievements are based on linguistic resources built at ATR Spoken Language Translation Research Laboratories (referred to as "ATR" hereafter). In this section, we briefly describe these resources and the three technical topics.

Figure 1.2 shows an overview of our thesis. The three technical topics are indicated in dark round boxes, and the linguistic resources used in this research are indicated in dashed boxes.

- Research Resources

  Our research is focused on the domain on the spoken language of travel conversation. Spoken language has interesting characteristics because it contains domain-specific expressions and more ungrammatical sentences than written text and these factors cause difficulty in spoken language processing.

  We use text corpora and MT systems created at ATR. The text corpora consist of

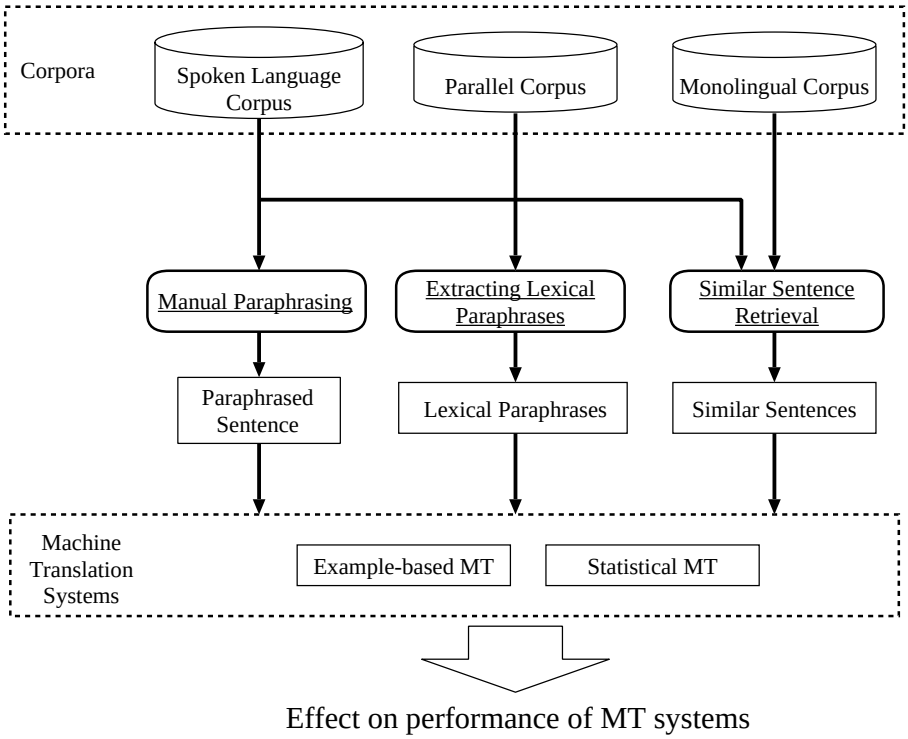

Figure 1.2. Thesis Overview

a transcribed speech text corpus[2] and a basic travel expression corpus. The former corpus is used for test sentences in experiments, and the latter is used for acquiring paraphrases. Two in-house MT systems are used for experiments: Example-based MT (EBMT) and Statistical MT (SMT). We test our paraphrasing methods on these two MT systems. Details of resources are described in Chapter 2.

- Analysis of Manual Paraphrases
  Before developing a method for acquiring paraphrases, we investigated human paraphrasing and clarified what types of paraphrases are dominant and how humans paraphrasing can be effective for MT (Shimohata et al., 2003a). These results are described in Chapter 3.

  Furthermore, we provided human paraphrased sentences to MT systems to verify

[2]The texts in the corpus are manually transcribed and free of speech recognition errors.

the effect of human paraphrasing (Shimohata et al., 2004a). These results are described in Chapter 4.

- Automatic Acquisition of Lexical Paraphrases
  Lexical paraphrases are the minimum unit for paraphrasing. Synonyms are typical lexical paraphrases. We develop a method for extracting lexical paraphrases from a parallel corpus that has two major advantages for MT: (1) it extracts function word related synonymous expressions that previous works have not considered, and (2) it extracts paraphrases not only from a monolingual viewpoint but also from a translingual viewpoint. Paraphrases in a corpus can be unified into a single expression by acquired paraphrases. Such paraphrase unification simplifies the corpus and improves the performance of corpus-based MT. We demonstrate the effect of paraphrase unification on EBMT and SMT systems. In addition, we carry out paraphrase acquisition from various language pairs and evaluated acquired paraphrases. Details are given in Chapter 5.

- Similar Sentence Retrieval
  We propose a method for retrieving a similar sentence from a monolingual corpus, which means sentential paraphrasing. This method supplements the lexical paraphrase acquisition method and has two advantages in that (1) it relies on a monolingual corpus that is easy to prepare, and (2) it acquires paraphrases whose differences go beyond the lexical level.

  We utilize this similar sentence retrieval technique for MT in two ways. One method is a pre-edit of an input sentence. If a given input sentence is too difficult to translate, a similar translatable sentence is retrieved and is given to the MT system instead of the original input sentence. The other is an example-based rough translation system. We use our similarity measure method with an EBMT system. An EBMT system using our retrieval method provides wider coverage of translatable sentences since the retrieval is robust. We named the translation method Rough Translation. Details are given in Chapters 6 and 7.

Chapter 2

Research Resources

Our research focuses on spoken language used in travel conversations. MT for spoken language has been developed in the field of speech-to-speech translation (S2ST). C-star,[1] Verbmobil (Wahlster, 2000), and Nespole! (Metze et al., 2002) are well-known projects in this field. S2ST technologies consist of speech recognition, MT, and speech synthesis (Waibel, 1996; Wahlster, 2000; Yamamoto, 2000). The MT part receives speech texts perceived by a speech recognizer. The characteristics of spoken language cause difficulty in the MT part since the styles of speech are different from those of written text and are sometimes ungrammatical (Lazzari, 2002). Therefore, rule-based MT cannot translate speech as accurately as written-style text.

ATR has been developing S2ST technology for fifteen years and has constructed many related resources. This research owes much to these resources. Recently, the laboratory has been enriching a multilingual corpus for training MT modules, since it adopts a corpus-based approach for MT development (Sumita et al., 2003).

1. Corpora

We use two types of corpora: a transcribed dialog text corpus (TDC) (Takezawa and Kikui, 2003) and a basic travel expression corpus (BTEC) (Kikui et al., 2003). TDC is used as a collection of input sentences for an S2ST system. BTEC is used as a learning corpus for our corpus-based MT systems.

[1] http://www.c-star.org/

Lang.	Sentence
Jpn.	ありがとうございますニューヨークシティホテルでございます.
Eng.	Hello New York city hotel.
Jpn.	もしもし部屋の予約をお願いしたんですが.
Eng.	I'd like to make a reservation for a room please.
Jpn.	かしこまりました.
Eng.	Sure.
Jpn.	失礼ですがお客様のお名前は.
Eng.	May I have your name.
Jpn.	はい田中弘子です.
Eng.	It's hiroko tanaka.

Figure 2.1. Sample Sentences from TDC

1.1 Transcribed Dialog Corpus

TDC data consists of transcribed speeches of simulated travel conversations , such as hotel reservations, conducted between two persons who speak different native languages. Either a human interpreter or a machine translation system assists the progress of dialogs. Dialogs were recorded, transcribed, and translated into the opposite language.

Sample utterances from TDC are shown in Figure 2.1. They are arranged in the order of the progress of the dialog, showing a characteristic of the data coming from its origin. Several sentences are concatenated into a single utterance, since a speech recognizer cannot recognize sentence boundaries inside an utterance. The only punctuation mark used is the symbol "." because a speech recognizer cannot recognize intonation.

ATR has developed two versions of TDC, SLDB and MAD, which differ in their translation facility. SLDB was constructed under a fully human-assisted translation environment, and MAD was constructed under a partially machine-aided translation environment. SLDB and MAD consist of 16,084 and 3,568 sentences, respectively. Table 2.1 shows the statistics of SLDB.

Table 2.1. Statistics of SLDB

	Japanese	English
Number of utterances	16,084	16,084
Number of sentences	21,769	22,928
Total number of words	236,066	181,263
Number of word entries	5,298	4,320
Average number of words/sentence	10.84	7.91

Table 2.2. Statistics of BTEC

	Japanese	English
Number of utterances	200,241	200.241
Number of sentences	220,244	223,482
Total number of words	1,689,442	1,230,650
Number of word entries	21,329	17,076
Average number of words/sentence	7.67	5.51

1.2 Basic Travel Expression Corpus

BTEC is a collection of edited colloquial travel expressions often found in phrasebooks. The corpus is a Japanese-English bilingual corpus. Table 2.2 shows the statistics of BTEC, and Figure 2.2 shows sample sentences from BTEC. Chinese and Korean translations were subsequently added to the corpus.

1.3 Difference of the Two Corpora

Unlike BTEC, the TDC corpus is derived from transcribed utterances. Although both corpora contain expressions frequently used in travel conversation, they have different characteristics.

The differences in average number of words and sentences are shown in Tables 2.1

Japanese Sentence	English Sentence
禁煙席をお願いします	A non–smoking seat, please.
喫煙席をお願いします	A smoking seat, please.
通路側の席をお願いします	An aisle seat, please.
窓側の席をお願いします	A window seat, please.
旅行の目的は何ですか	What is the purpose of your visit?
何日滞在の予定ですか	How long are you going to stay here?
帰りの航空券を持っていますか	Do you have a return ticket?
所持金はいくらですか	How much money do you have with you?
宿泊先はどこですか	Where will you be staying?
この鞄を開けて下さい	Please open this bag.

Figure 2.2. Sample Sentences from BTEC

Table 2.3. Cross Perplexity

		Language Model	
		BTEC	TDC
Test	BTEC	16.4	58.3
	TDC	72.3	16.3

and 2.2. Sentences in TDC are more than two words longer than those in BTEC, in both English and Japanese, because sentences in TDC contain unnecessary words or subordinate clauses, which assist the listener's comprehension and avoid the possibility of seeming rude.

Table 2.3 shows the cross perplexity between BTEC and TDC (Takezawa et al., 2002). Perplexity functions as a metric for how well a language model derived from a training set matches a test set (Jurafsky and Martin, 2000). The cross perplexities between BTEC and TDC are much higher than the self-perplexity of either of the two corpora. This result also illustrates the great difference between the two corpora.

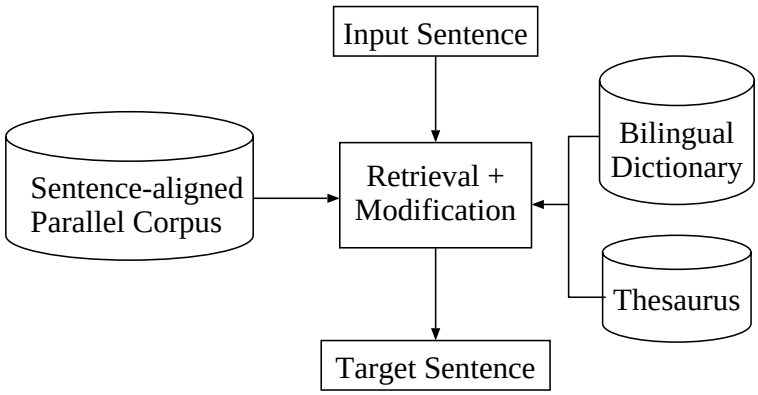

Figure 2.3. Structure of EBMT System

2. Corpus-based Machine Translation Systems

2.1 Example-based Machine Translation System

EBMT is a promising method for S2ST because it enables robust translations of ungrammatical sentences and requires far less manual work than rule-based MT. Its basic idea is retrieval of sentences similar to input sentences from a parallel corpus and modification of the translation of similar sentences to generate output translation (Nagao, 1981). Figure 2.3 shows the structure of our EBMT system. EBMT outputs no translation if there is no similar example sentence in the corpus. It uses the BTEC corpus as an example corpus.

The translation process consists of the following four steps.

1. Retrieve the most similar translation pair
   In this process, the most similar sentence is retrieved from the example corpus. The similarity is measured by the edit distance between word sequences of the input and example sentences. The weight of substitution is adjusted for the similarity of two words, based on the hierarchical distance in the given thesaurus. An example sentence that has larger similarity than a predefined threshold is discarded.

2. Generate translation patterns

Translation patterns are generated from the retrieved translation pairs. Corresponding words between translation pairs are detected by consulting a bilingual dictionary. Then, these corresponding words are replaced with variables to create translation patterns.

3. Select the best translation pattern
   If multiple translation patterns are obtained, we have to select the most suitable one according to the following conditions, listed in order of priority.
   (a) The translation pattern with the highest frequency.
   (b) The translation pattern with the highest frequency of entire words in the pattern.

4. Substitute target words for source words
   A target sentence is generated by converting the variables in the target part of the selected translation pattern into target words.

2.2 Statistical Machine Translation System

The framework of statistical machine translation formulates the problem of translating a sentence from language J into another language E as the maximization problem of the conditional probability $\hat{E} = argmax_E P(E|J)$ (Brown et al., 1990).

The application of the Bayes Rule results in $\hat{E} = argmax_E P(E)P(J|E)$. The former term $P(E)$ is called the language model, representing the likelihood of E. The latter term $P(J|E)$ is called the translation model, representing the generation probability from E into J. Parameters in both models are automatically determined from learning data. This SMT system uses BTEC as learning data.

A decoder, which generates target sentences, plays an important role in SMT. This SMT system adopts an innovative decoding strategy: example-based decoding (Watanabe and Sumita, 2003). In this strategy, the most similar sentence is retrieved from an example corpus. Then, its translation is modified several times in looking for the sentence with the highest probability.

This method resembles EBMT because it utilizes translation of the most similar sentence. It outperforms a traditional decoding method that decodes word-by-word and generates an output string word-by-word.

Chapter 3

Analysis of Manual Paraphrasing

Natural language has a wide variety of ways to express identical meaning. Humans can subconsciously use and recognize a variety of paraphrases. We can consider many types of paraphrasing. For example, "Please open the window" can be paraphrased in many ways. With a small change we can paraphrase it as "Would you open the window?" making it more polite than the original. We can convey this request by a completely different sentence: "I feel hot and need a breeze." This phenomenon poses a question: What are the characteristics of human paraphrasing? Answering this question is crucial to overcoming the difficulty caused by human paraphrasing.

In this chapter, we report on investigations into human paraphrasing. First, we classify paraphrasing ranges into three types: sentential, phrasal, and lexical. Then, two human paraphrasers receive sentences to be paraphrased (called "original sentences" hereafter) and make paraphrased sentences by using these three paraphrasing methods. We report the degree of variation by paraphrasing type as well as individual variation.

1. Paraphrasing Types

We classify paraphrases into three types: sentential, phrasal, and lexical, based on the range of difference between the original and paraphrased sentences. A brief definition of each type is given below.

Sentential: Change in the principal part of a sentence, such as subject, predicate, or modality.

Original 1	Can I reserve some seats?
Paraphrased 1-1	I'd like to reserve some seats.
Paraphrased 1-2	Can I make a reservation for some seats?
Original 2	Do you have any vacant seats?
Paraphrased 2-1	Is there a vacant seat?
Paraphrased 2-2	I was wondering if there were any vacant seats.

Figure 3.1. Examples of Sentential Paraphrasing

Phrasal: Syntactical changes that extends beyond a phrase.

Lexical: Lexical changes within a phrase.

In this classification, each paraphrasing type requires different knowledge to deal with. Sentential paraphrasing requires pragmatic knowledge, phrasal paraphrasing requires syntactic knowledge, and lexical paraphrasing requires lexical knowledge. Therefore, the ratio of each paraphrasing type is very useful in assigning development costs to each type.

1.1 Sentential Paraphrase

Sentential paraphrasing changes the principal information of the original sentence. Principal information refers to the headword of the subject, the predicate, and modality. When headwords are changed, if the changed words are semantically equivalent, this change goes to lexical paraphrasing: for example, "speak" and "talk."

Examples of sentential paraphrases are shown in Figure 3.1. Sentences 1 and 1-1 differ in modality. Sentence 1 shows a "question" modality while sentence 1-1 represents "request." Sentences 1 and 1-2 differ in the headword of the predicate, although they share a common subject. Sentence 1 has "reserve" as the headword of the predicate while sentence 1-1 has "make."

Original 1	The repair cost is ninety-nine dollars.
Paraphrased 1-1	The cost of repair is ninety-nine dollars.
Original 2	Can I have a refund for my ticket?
Paraphrased 2-1	Can I have my ticket refunded?

Figure 3.2. Examples of Phrasal Paraphrasing

1.2 Phrasal Paraphrase

Phrasal paraphrasing does not satisfy the conditions for sentential paraphrasing, but it changes the structure of a phrase or order of phrase elements. Figure 3.2 shows examples of phrasal paraphrasing. Sentences 1 and 1-1 differ in the structure of the subject phrase. Sentences 1 and 2-1 differ in the order of objects.

1.3 Lexical Paraphrase

Lexical paraphrasing simply replaces words, and does not satisfy the conditions of either sentential or phrasal paraphrasing, i.e. does not change structure of the original sentence.

Variety of lexical paraphrasing is made by using brackets, parentheses, and the symbol "|." The expression "[A|B]" denotes the word "A" or the word "B." The expression "(A|B)" denotes the word "A," the word "B," or null-string[1]. Brackets and parentheses can be nested.

Figure 3.3 shows examples of lexical paraphrasing. Sentences 1 and 1-1 have different predicates. These verbs have similar meaning in the context of the original sentence. Sentences 2 and 2-1 have different auxiliary verbs. Paraphrased sentences 1-1 and 2-1 each produce five paraphrased sentences[2] by expanding the lexical variations.

[1] In other words, the expression can be omitted.

[2] A paraphrased sentence identical to the original sentence was excluded.

Original 1	You will see it on this side of the road.
Paraphrased 1-1	You will [see|find] it on this side (of the [road|street]).
Expanded 1-1	You will see it on this side of the street.
Expanded 1-2	You will see it on this side.
Expanded 1-3	You will find it on this side of the road.
Expanded 1-4	You will find it on this side of the street.
Expanded 1-5	You will find it on this side.
Original 2	May I make a call?
Paraphrased 2-1	[May | Can | Could] I make a (phone) call?
Expanded 2-1	May I make a phone call?
Expanded 2-2	Can I make a call?
Expanded 2-3	Can I make a phone call?
Expanded 2-4	Could I make a call?
Expanded 2-5	Could I make a phone call?

Figure 3.3. Examples of Lexical Paraphrasing

2. Construction of Paraphrase Corpus

Paraphrased sentences were written by two human paraphrasers who speak different varieties of English: Paraphraser A is an American English native, and Paraphraser B is a British English native. We randomly extracted 1,009 sentences from BTEC and provided them to the paraphrasers. These original sentences average 6.2 words. Each paraphraser independently generated paraphrased sentences according to the three paraphrasing types.

If a paraphraser found it difficult to paraphrase a certain sentence, he skipped it. As a result, paraphraser A paraphrased 983 original sentences and paraphraser B 996 sentences. A total of 970 sentences were paraphrased by both paraphrasers.

We told the paraphrasers to not use the following types of paraphrasing, which we excluded:

1. Different spelling (theater $\Rightarrow$ theatre).

Original Sentence
I would like to ask something of you.

Paraphraser A		
Sentential	Phrasal	Paraphrased Sentence
1	1	I [would like | want] to ask something of you.
1	2	I [would like | want] to ask you something.
2	1	[Could | Can] I ask something of you?
2	2	[Could | Can] I ask you something?

Paraphraser B		
Sentential	Phrasal	Paraphrased Sentence
1	1	I [would like | want] to ask something of you.
1	2	I [would like | want] to ask you to do something.
2	1	I have a favor to ask?
3	1	[Can | Could | Will | Would] you do me a favor?
3	2	[Can | Could | Will | Would] you do something for me?
4	1	I was wondering if you could (possibly) do me a favor?
4	2	I was wondering if you could (possibly) do something for me?
5	1	You could not do me a favor, could you?
5	2	You could not do something for me, could you?

Figure 3.4. Examples of Paraphrased Sentences

2. Paraphrasing to abbreviated spelling (I have ⇒ I've).

3. Exchange of concrete object and pronoun
   (Show your passport ⇒ Show this).

4. Change of word order.

Figure 3.4 shows an example of paraphrased sentences. In the figure, the numbers in the "Sentential" and "Phrasal" fields denote ID numbers for each paraphrasing type.

Table 3.1. Number of Acquired Paraphrased Sentences by Each Type

# of Sentences	Paraphraser A	Paraphraser B
Original Sentences	970	970
Sentential	3,971	4,314
Phrasal	4,387	5,961
Lexical	64,897	50,096

3. Analysis of Paraphrased Sentences

Table 3.1 shows the number of paraphrased sentences for each paraphrasing type. Paraphraser A made 3,971 sentences from the original 970 sentences by sentential paraphrasing. Then, these paraphrased sentences were expanded to 4,387 by phrasal paraphrasing. Then, lexical variations were embedded in these sentences. The expansion of lexical variations resulted in the generation of a total of 64,879 paraphrased sentences. Paraphraser B made 4,314, 5,961, and 50,096 sentences by sentential, phrasal, and lexical types, respectively.

An expansion ratio can be determined from the results. Table 3.2 shows expansion ratios by type and the totals. The expansion ratio of lexical paraphrasing is the highest among the three types for the two paraphrasers. Expansion ratios are larger in order of lexical, sentential, and phrasal.

Next, let us examine the difference between the two paraphrasers. Figure 3.5 illustrates the generated paraphrased sentences by each paraphraser and their overlap. Paraphraser A generated 64,879 paraphrased sentences, and Paraphraser B generated 50,096 sentences. By A, the ratio of the overlap to the sentences is 17.0%, and by B it is 15.05%. We believe this small ratio of overlap reflects the difference between American English and British English.

Table 3.2. Expansion Ratio for Each Paraphrasing Type

Expansion Ratio	A	B	Geometric Mean
Sentential	4.09	4.45	4.25
Phrasal	1.10	1.38	1.23
Lexical	14.79	8.40	11.15
Total	68.14	53.46	60.36

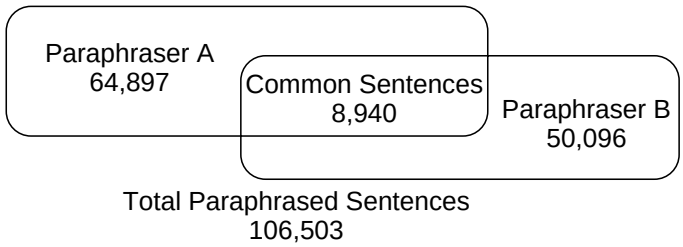

Figure 3.5. Overlap of Paraphrased Sentences between Paraphrasers

4. Related Work

Sugaya (Sugaya et al., 2002) and Kinjo (Kinjo et al., 2003) investigated how Japanese translations can be generated from an English sentence. They reported that, on average, an English sentence can produce 2,747 Japanese translations. This translation variety is much larger than our expansion ratio of 60.36. However, there is a great difference in the paraphrasing condition between their study and our study. Several paraphrasers collaborated in generating the translations in Sugaya's work, while only two paraphrasers individually did the paraphrasing in our work. Therefore, Sugaya's expansion ratio represents exhaustive paraphrasing that aims to encompass every possible variation. On the other hand, our expansion ratio represents individual paraphrasing that aims to provide practical variations. The difference in language, Japanese in Sugaya's work and English in our work, should also be considered for the difference in expansion ratios.

5. Conclusion

We analyzed human paraphrasing. Two human paraphrasers received original sentences extracted from transcribed speech texts and paraphrased them according to three paraphrasing methods: lexical, phrasal, and sentential. These three methods differ in range of paraphrased parts. Information on these paraphrasing types is useful for the effective development of MT systems since each type requires different methods.

An analysis of acquired paraphrased sentences clarified that lexical paraphrasing has the highest expansion ratio followed by sentential and phrasal paraphrasing. This tendency appeared with two paraphrasers. The average expansion ratios of lexical, sentential, and phrasal paraphrasing were 11.15, 4.25, and 1.23, respectively. In total, the average expansion ratio reached 60.36.

In addition, we found that the two paraphrasers, an American English native and a British English native, had a small overlap of their paraphrased sentences. This seems to reflect the differences in their native languages. However, the expansion ratios of both paraphrasers for each paraphrasing type are similar.

Chapter 4

Building a Paraphrase Corpus for Speech Translation

When an MT system receives input sentences of spoken language, the following two types of sentences are difficult to translate: (1) long sentences and (2) sentences having redundant expressions, which often exist in spoken language. To reduce these difficulties, we developed methods to paraphrase input sentences into more translatable ones.

Here, we describe a Japanese paraphrase corpus to clarify the effect of paraphrasing on speech translation. The corpus consists of original sentences derived from travel conversation and versions of them paraphrased by humans. We use three paraphrasing methods: plain, segment, and summary paraphrasing. Plain paraphrasing is applied to short sentences, where redundant expressions are replaced with plain ones. Segment and summary paraphrasing are applied to long sentences, where long sentences are converted into one or several short sentences. Then, we report a comparison of machine translation quality between the original sentences and the paraphrased sentences. We also report the statistical characteristics of these sentences in terms of perplexity.

1. Basic Idea of Paraphrasing

We focused our paraphrasing on the following types of sentences, since they often cause degradation of translation quality.

1. Long input sentences
   In general, the longer input sentences become, the worse the MT quality is. This is because long sentences have many candidate translations, and it is difficult to select the appropriate one among them.
   To reduce this disadvantage, we use paraphrasing methods that paraphrase long sentences into one (**summary**) or several (**segment**) short sentences.

2. Input sentences with redundant expressions
   Redundant expressions are often found in spoken language. These expressions have the effects of assisting the listener's comprehension and avoiding the possibility of giving the listener a curt impression. On the other hand, they lengthen the sentences and cause translation errors.
   To reduce this disadvantage, we use paraphrasing methods that replace redundant expressions with plain ones (**plain**).

In our paraphrasing strategy, it is important to classify input sentences into short or long. We describe the adopted metric of sentence length and the threshold used to determine short or long in the following sections.

1.1 Sentence Length Metric

We use "number of content words" as a metric of sentence length. This means that the unit of the metric is a word, and function words are excluded from the word count. The reason for excluding function words is that they have a wide variety in Japanese conversation. This variety reduces the correlation between the number of function words and the complexity of the sentences. Moreover, shortening sentences by deleting function words sometimes causes translation difficulty because an MT system has to infer the lost function word information.

Content words and function words are classified by information on part-of-speech. Content words are defined to include nouns, verbs, adjectives, adverbs, and numerals. Function words are defined to include particles, auxiliary verbs, and the copula. Compound words, for example, "New York," "get off," and "two hundred dollars," are treated as one word in the case of English.

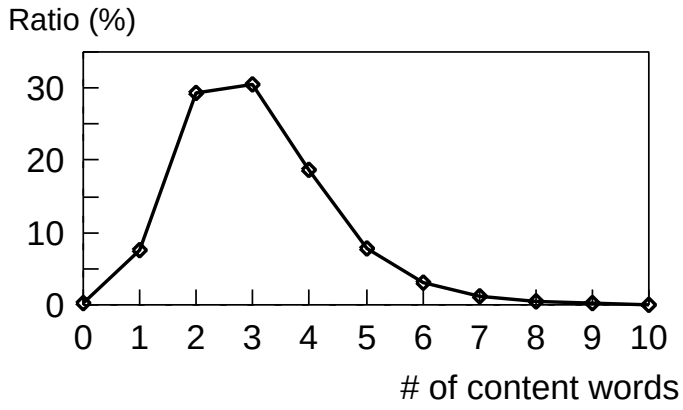

Figure 4.1. Ratio of Content Words in BTEC Corpus

1.2 Threshold for Long Sentences

We define "short" sentences as sentences having less than five content words. This threshold is based on statistics of the BTEC corpus (Kikui et al., 2003), which is a fundamental bilingual corpus for developing our corpus-based MT systems. In the BTEC corpus, sentences having fewer than five content words are dominant (86.5%) over those having five or more content words. Figure 4.1 shows the ratio of content words in the BTEC corpus.

2. Paraphrasing Methods

In this section, we describe the details of three paraphrasing methods. Examples of input sentences and paraphrased sentences are shown in Figure 4.2. To facilitate the reader's understanding, the examples are shown in English and content words are underlined.

2.1 Segment Paraphrasing

A long sentence is divided into several short sentences in the segment paraphrasing method. Some added words are allowed if needed to make complete sentences. If it

Org.	the twin room facing the ocean is three hundred dollars per night
Seg.	there is a twin room facing the ocean.
	it is three hundred dollars per night
Org.	Let me just check my computer and get back to you on that.
Sum.	I will check and get back to you on that.
Org.	I was hoping you'd have a triple room.
Plain	I'd like a triple room.

Figure 4.2. Examples of Paraphrasing

is difficult to divide a given sentence into parts that are all "sentences," dividing the sentence into "phrases" is allowed. In the first example in Figure 4.2, the original sentence includes five content words. It is paraphrased into two short sentences, each of which includes fewer than five content words.

2.2 Summary Paraphrasing

A long sentence is condensed into one short sentence by eliminating unimportant content words in the summary paraphrasing method. We assume that a good translation of condensed sentences is more valuable than a poor translation of original sentences. In the second example in Figure 4.2, the number of content words is reduced from five to three. Deleted information such as "just" and "computer" is considered insignificant.

2.3 Plain Paraphrasing

Redundant expressions in input sentences are replaced by plainer ones in the plain paraphrasing method. Furthermore, insignificant information can be deleted. Insignificant information is defined as information that can be removed without causing a significant problem for the progress of the conversation. We leave the judgment of redundant and plain expressions to a human paraphraser. In the third example in Figure 4.2, the original sentence includes euphemistic and polite expressions, while the paraphrased sentence is a plain one. This paraphrasing strategy is also applied to segment paraphrasing and summary paraphrasing.

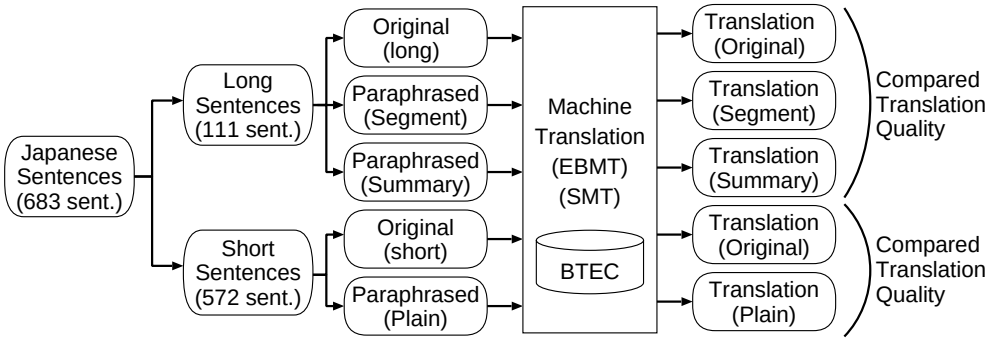

Figure 4.3. Overview of Experiment

3. Experiment

We built a Japanese paraphrase corpus consisting of 683 original sentences and corresponding paraphrased sentences for a pilot study. The original sentences were derived from the travel conversation corpus (Kikui et al., 2003).

Figure 4.3 shows an overview of the experiment. We classified input sentences into short and long and had a human paraphraser paraphrase them. Short sentences have one paraphrased sentence of plain paraphrasing, and long sentences have two paraphrased sentences of segment and summary paraphrasing. This paraphrasing task took just one day for each method.

We gave the original sentences and paraphrased sentences to MT systems and obtained translations. The effect of paraphrasing on MT was evaluated by comparing the translation qualities of original and paraphrased sentences (Section 3.2). The characteristics of original and paraphrased sentences were analyzed by using perplexity (Section 3.3).

3.1 Experimental MT Systems

Two corpus-based MT systems were used in the experiment: Example-based MT (EBMT) (Sumita, 2001) and Statistical MT (SMT) (Watanabe and Sumita, 2003). The two MT systems commonly use the BTEC corpus as an example/learning corpus.

The basic idea of the EBMT system is that it retrieves sentences similar to input

Table 4.1. Translation Quality

Sent. Length	Trans. Method	Method	Ratio of Evaluation (%)					Trans. Acc. (%)
			A	B	C	D	N	
Long	EBMT	Original	7.2%	14.4%	10.8%	5.4%	62.2%	32.4%
		Segment	20.7%	28.8%	20.7%	29.7%	0.0%	70.2%
		Summary	17.1%	17.1%	14.4%	20.7%	30.6%	48.6%
	SMT	Original	21.6%	22.5%	16.2%	39.6%	-	60.3%
		Segment	27.9%	20.7%	12.6%	38.7%	-	61.2%
		Summary	20.7%	14.4%	22.5%	42.3%	-	57.6%
Short	EBMT	Original	67.0%	9.8%	2.6%	11.0%	9.6%	79.4%
		Plain	66.8%	12.6%	3.3%	11.0%	6.3%	82.7%
	SMT	Original	68.2%	11.2%	5.4%	15.2%	-	84.8%
		Plain	69.2%	10.5%	5.2%	15.0%	-	85.0%

sentences from a parallel corpus and modifies the translation of similar sentences to generate the output translation. The similarity between input sentence and example sentences is measured by edit distance. The weight of substitution is adjusted by similarity, which is based on the given thesaurus.

The basic idea of the SMT system is that it generates output translation that has the highest likelihood for the input sentence. Likelihood is decomposed into a translation model and a language model. The parameters for the two models are determined from a learning corpus.

3.2 Translation Quality

The MT systems receive both original and paraphrased sentences and return their English translations. These translations are evaluated by a native English speaker. There are four evaluation ranks: A (good), B (fair), C (acceptable), and D (bad). The EBMT system outputs no translation when there is no similar sentence in the example corpus. In this case, we give an "N" rank.

Table 4.1 shows the results of translation quality. Translation accuracy is defined as the ratio of sentences having A, B, and C ranks to total sentences. As for long sentences, both paraphrasing methods provide a large improvement in EBMT. In par-

Table 4.2. Cross-perplexity

Test Data	Cross-perplexity
Original (long)	61.7
Segment	39.3
Summary	45.8
Original (Short)	32.7
Plain	24.7

ticular, the paraphrasing methods improve performance by reducing the ratio of the N rank. Segment paraphrasing reduces it from 62.2% to 0.0%, and summary reduces it to 30.6%. On the other hand, both paraphrasing methods bring little improvement in SMT. As for short sentences, the ratios of all ranks are approximately equal. This shows that plain paraphrasing for short sentences has little effect.

3.3 Cross Perplexity

Cross perplexity (CP) is a metrics for determining how predictive the test data is when using the N-gram model learned from training data. The lower the CP value, the more predictive the test data is with the learning data. The CP in which the BTEC corpus is used as training data indicates dissimilarity between the BTEC corpus and the test data.

Table 4.2 shows the CP value using the original and paraphrased sentences as test data. All CP values of the paraphrased sentences are lower than those of the original sentences. This effect is more evident in long sentences, and it indicates that all paraphrasing methods simplify the original sentences and make the paraphrased sentences more predictive for the BTEC corpus.

Table 4.3. Positive/Negative Paraphrasing Cases

Sentence Length	Translation Method	Paraphrasing Method	Rank Comparison with Original		
			Para.>Org.	Para.=Org.	Para.<Org.
Long	EBMT	Segment	77.5%	9.9%	12.6%
		Summary	44.1%	44.1%	11.7%
	SMT	Segment	30.6%	45.0%	24.3%
		Summary	19.8%	51.4%	28.8%
Short	EBMT	Plain	9.3%	83.7%	7.0%
	SMT	Plain	8.7%	83.7%	7.5%

3.4 Positive/Negative Paraphrasing between Original and Paraphrased Sentences

In this section, we discuss the change in MT quality in detail. Table 4.3 shows the ratios of positive/negative cases for both the original and paraphrased sentences. Here, "positive" means the paraphrased sentence has higher-ranked MT quality than the original sentence, while "negative" means the opposite case. The comparison of MT quality between paraphrased (Para.) and original (Org.) sentences is based on the ranks of A(Good), B(Fair), C(Acceptable), D(Non-sense), and N(No-translation).

The results show that the ratios of positive/negative cases differ between long and short sentences. Nearly half of the paraphrased sentences have a different evaluation rank from that of the original sentences for long sentences, while almost all paraphrased sentences remain at the rank of the original sentences for short sentences. The paraphrasing effect for long sentences depends on the occurrence of negative cases. The EBMT system had relatively few negative cases and showed large improvement. However, the SMT system had many negative cases and showed little improvement.

This result suggests that the paraphrasing effect can be improved by eliminating negative paraphrasing. We consider the following two works useful for this elimination. (Shimohata and Sumita, 2002) proposed a method for extracting local paraphrases from two sentences sharing the same meaning. We can obtain local paraphrases by applying this method to the original and paraphrased sentences. (Imamura et al., 2003) proposed

a filtering method of translation rules by automatic evaluation of machine translation. Local paraphrases that are effective for target MT systems can be filtered by this method. The combined use of these approaches remains our future work.

4. Conclusions

We are developing a method of paraphrasing input sentences to facilitate machine translation. In this chapter, we reported a Japanese paraphrase corpus. The corpus consists of original sentences derived from a travel conversation corpus and their paraphrased versions. We used three paraphrasing methods: plain, segment, and summary. Plain paraphrasing is applied to short sentences and replaces redundant expressions with plainer ones. Segment and summary paraphrasing are applied to long sentences to convert them into one of several short sentences.

Experimental results suggest that this paraphrasing strategy has a large effect on EBMT for long sentences but a small effect on SMT for long sentences; in addition, paraphrasing has a small effect on both MTs for short sentences. We believe that additional improvement can be achieved by eliminating deteriorating paraphrasing.

At present, we are constructing a paraphrased corpus containing about forty-five thousand sentences in both Japanese and English. We plan to exploit this corpus and thus improve the effect of our paraphrasing in both Japanese and English.

Chapter 5

Extracting Lexical Paraphrases from a Parallel Corpus

This chapter focuses on extraction of lexical paraphrases and their application to MT systems. We call lexical paraphrases as "synonymous expressions" (SE).

A variety of SE causes difficulties in NLP applications such as machine translation and information retrieval. Let us consider the problems in example-based machine translation (EBMT). The basic idea of EBMT is to retrieve a similar sentence from an example bilingual corpus and to modify the translation of the retrieved sentence to obtain an output translation (Nagao, 1981). When it measures the similarity between an input sentence and sentences in the example corpus, variations in SE prevent similar sentence retrieval since SE degenerate superficial similarity, increasing the distance between essentially similar sentences and making them appear dissimilar.

Next, let us look at the case of statistical machine translation (SMT), which is a translation method based on a noisy channel model (Brown et al., 1990). It carries out translation by generating a sentence in the target language that has the highest likelihood for a given input sentence. Parameters for calculating likelihood are determined from a learning corpus. Variations in SE complicate both language and translation models and cause degradation in translation performance due to data sparseness.

The proposed SE extraction method has the advantage that it requires only a tagged parallel corpus. The extracted SE feature the following three points.

1. They include expressions that are synonymous from the translingual viewpoint.

2. They consist of synonyms as well as surrounding words as contextual information.

3. They include many function word related synonyms.

SE extraction is based on substitution of edit-operations between synonymous sentences (SS) that share the same translation. In addition, frequency based filtering refines the extracted SE. This SE extraction is iterated by unifying the previously extracted SE in both languages.

We demonstrate the effect of unifying extracted SE on improving the performance of the above two corpus-based machine translation methods, i.e., EBMT and SMT. In an experiment, a corpus is self-unified by using its expression variety. Then, the original corpus and the self-unified corpus are used for a comparative experiment.

We describe the features of our SE in Section 1, the extraction method in Section 2, and the experiments on EBMT and SMT in Section 3. We describe related research and the advantage of our method in Section 4.

Then, we describe comparative experiments conducted to investigate the influence of applied language pairs on the accuracy of extracted paraphrases in Section 5.

1. Features of Synonymous Expressions

1.1 Synonymous from the Translingual Viewpoint

In general, "*synonymous*" refers only to the "monolingual" viewpoint. When we utilize synonyms for machine translation, synonyms from not only the monolingual viewpoint but also the translingual viewpoint are useful.

Figure 5.1 shows an example of synonyms from the translingual viewpoint[1]. The two sentences in example (1) differ in the words 'wallet" and "purse." Although this difference is significant and these words are not considered equivalent in English-language culture, the two words are translated into the same Japanese word "saifu." Therefore, these two words can be equated when translating into Japanese. The sentences in example (2) show a similar case for Japanese-to-English translation. The two sentences show a similar situation related to the Japanese words "ane" (older sister) and "imouto" (younger sister). Although the difference cannot be equated in Japanese culture, they

[1]Japanese sentences are segmented into words by spaces.

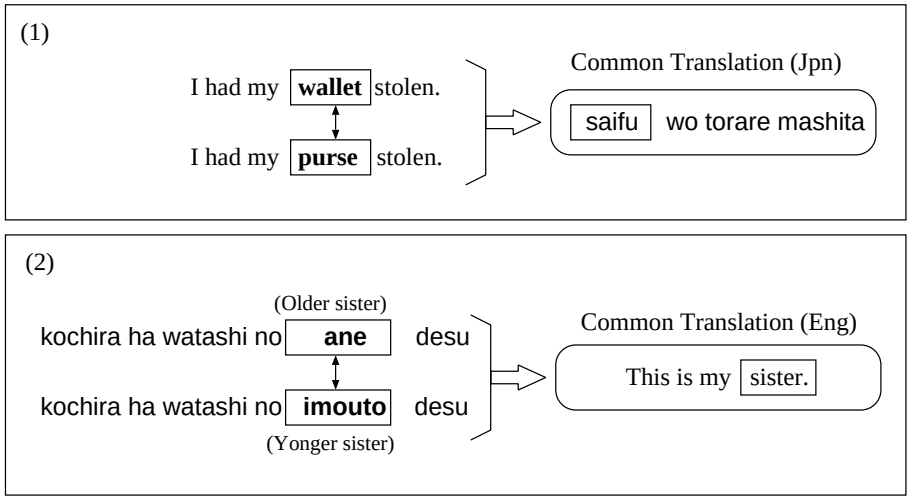

Figure 5.1. Synonyms from the Translingual Viewpoint

are translated into the same English word "sister" since it is not usual to express age relation in English-language culture. In another case, the singular/plural difference of nouns is insignificant in Japanese since the number is rarely expressed in a Japanese sentence.

When considering synonyms for translation, these translingual differences are useful. Our method extracts synonyms from the translingual viewpoint, since synonymy is based on the equivalence of their translations.

1.2 Expressions Including Context

The synonymy of two words often depends on their surrounding context. For example, let us consider the synonymy of the two words "photos" and "pictures." The word "photos" has only the meaning of *photographs*, while the word "pictures" has the two meanings of *photographs* and *paintings*. The synonymy of the two words cannot be determined without the context since the meaning of picture is ambiguous. This ambiguity can be solved by considering the words surrounding it. If "picture" is embedded in the expression "take pictures of," it denotes *photos*. In "draw pictures of", it denotes *paintings*. This example illustrates that we need to take the surrounding words of two

words into account when determining their synonymy. Hereafter, we define an "expression" as a word sequence that consists of not only synonyms but also their surrounding words.

There are several ways to take in surrounding words in terms of direction and length. We adopt the setting of a "single" word from "both" sides of synonyms for all languages and all types of synonyms. Our preliminary experiment proves that the setting of taking in two words from both sides greatly reduces the amount of extracted SE and does not improve the precision of SE compared with the condition of a single surrounding word on both sides.

1.3 Function Word Related Expressions

Our method has the advantage that it can extract synonyms related not only to content words but also to function words. Function word related synonyms are rarely described in existing thesauri and are difficult to acquire automatically from corpora (described in Section 4). However, they have a wide variety and are especially important in conversation. For example, there are various expressions for expressing a request such as "Would you ...," "Could you ..," "Would you mind if I ....," and "Please ...."

Our method can extract many function word related SE owing to two features: the same translation assures the synonymy of SE (described in Section 1.1) and the inclusion of contextual information increases the amount of extracted paraphrases (described in Section 1.2).

1.4 Lexical Synonymous Expressions

We focus our target SE on lexical expressions, where the synonym parts are confined to phrases. Figure 5.2 shows examples of lexical and non-lexical differences. In the figure, the differences from the original sentence are written in bold. Brackets enclose lexical differences and parentheses enclose non-lexical differences.

We define a lexical SE such that it contains only one different part in it. For example, the two sentences "**Are** you **taking** any pills regularly?" and "**Do** you **take** any pills regularly?" have two different parts. These two different parts are related to each other and we cannot handle either of them separately. We regard these independent differences as a non-lexical difference. According to this definition, the difference of

Original Sentence	Will you show me your credit card?
Synonymous Sentence 1	[**Would**] you show me your credit card?
Synonymous Sentence 2	[**Could**] you show me your [**card**]?
Synonymous Sentence 3	(**May I see**) your credit card?
Synonymous Sentence 4	(**I would like to see**) your credit card.

Figure 5.2. Examples of Lexical Difference

word order change is defined as non-lexical. The difference between the two sentences "Would you **please** call me a taxi?" and "Would you call me a taxi **please**?" is an example of word order change.

The criterion for judging whether an extracted expression is lexical is the number of words in the synonym part, and thus it does not depend on syntactic information. We defined lexical expressions as those having a difference within two words after a preliminary experiment.

2. Extraction of Synonymous Expressions

SE extraction is carried out by iterating the basic extraction procedure. We describe the basic extraction procedure in Section 2.1 and the iteration in Section 2.2.

In this section, we use English SE extraction for our explanation. We represent a parallel corpus as a collection of English and Japanese sentence pairs, such as $\{(Js_1, Es_1)$, (Js_2, Es_2), ... , $(Js_n, Es_n)\}$.

2.1 Basic Extraction Process

Grouping Synonymous Sentences

SS groups are formed by gathering sentences that share the same translation. If Japanese sentences Js_1, Js_4, and Js_{11} are the same, English sentences $\{Es_1, Es_4, Es_{11}\}$ form a single SS group. Figure 5.3 shows an example of an English SS group in which sentences share the same Japanese translation "shashin wo totte mo ii desu ka."

Common Translation (Jpn)	shashin wo tottemo iidesuka
Synonymous Sentences (Eng)	(1) Can I take pictures? (2) May I take photos? (3) May I take some photos? (4) Can I take a photo? (5) Is it OK to take pictures?

Figure 5.3. Synonymous Sentence Group

Extraction of Synonymous Expression Pairs

First, all combinations of SS pairs are extracted from an SS group. For example, an SS group containing five sentences provides ten SS pairs.

Second, edit-operations between each SS pair are extracted as follows:

1. Apply DP-matching (Cormen et al., 2001) to an SS pair, regarding sentences as word-sequences sandwiched by "#" (start-of-sentence) and "%" (end-of-sentence). Words are identified by their surface forms and POSs.

2. An SS pair having more than two edit-distances is removed from consideration. Each weight of the edit-operation, such as insertion, deletion, and substitution, is counted as 1. This filtering is effective to avoid SS pairs having non-lexical SE.

3. Edit-operations of substitution and its surrounding words are extracted. We exclude insertion and deletion operations from extraction since most of them are non-lexical, mainly involving with word order change.

Figure 5.4 shows extracted SE pairs from SS pairs of sentences (1) and (2) in Figure 5.3. DP-matching between the two sentences proves that these sentences have two edit-distances consisting of the substitution of "can" and "may" and the substitution of "photos" and "pictures." This SS pair is valid since it has two edit-distances. Two SE pairs, as shown in the lower side of Figure 5.4, are acquired by extracting the substitution of the edit-operation and its surrounding words.

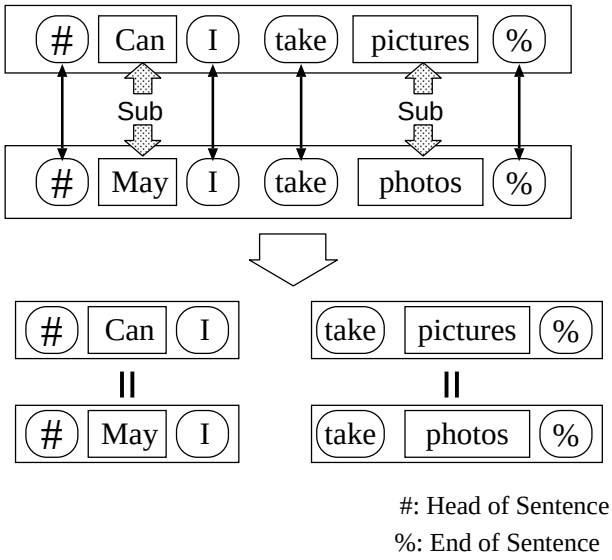

Figure 5.4. Extraction of Synonymous Expression Pairs

Filtering Synonymous Expression Pairs

The extracted SE pairs are filtered by two frequency-based conditions. The thresholds are determined experimentally.

Ratio of SE pair frequency to component expression frequencies

"Component expressions" of the SE pair "Exp1 = Exp2" denote Exp1 and Exp2. The SE pair "Exp1 = Exp2" is tested by comparing the ratio of the SE pair frequency [2] with that of each component expression Exp1 and Exp2. This filters out SE pairs whose synonymy is true under a certain situation. We set the condition for this filtering as follows.

$$\frac{freq(\text{“}Exp1 = Exp2\text{”})}{min(freq(\text{“}Exp1\text{”}), freq(\text{“}Exp2\text{”}))} > 0.05$$

Frequency of SE pairs

SE pairs appearing only once are filtered out. This condition is effective for removing improper SE pairs containing a free translation or spelling error.

[2]Frequency of expression means the number including the SS group.

Clustering Synonymous Expressions

The preceding processes extract SE pairs from a parallel corpus. These extracted SE "pairs" are clustered into "SE clusters" based on the transitive relation. If SE pairs {Exp1 = Exp2} and {Exp2 = Exp3} are extracted, an SE cluster {Exp1, Exp2, Exp3} is formed.

Next, the most frequent expression in an SE cluster is selected as the "representative expression." We consider that the most frequent expression to be the one most common or applicable to the broadest range of situations. Extracted SE clusters can simplify a given text by replacing a non-representative expression with the representative expression.

2.2 Iteration

The basic extraction procedure described in Section 2.1 extracts English SE clusters. The procedure can also extract Japanese SE clusters by reversing the roles of English sentences and Japanese sentences. Unifying SE by extracted SE clusters in both languages can integrate some SS groups that were separated in the previous status. Additional SE clusters can be acquired by repeating the basic extraction procedure after SE unification.

Figure 5.5 shows an example of integrating two SS groups by unification. In the upper-left side of the figure, two English SS groups are formed by the first basic extraction. The translations of the two SS groups have the difference of "kuremasenka" and "kudasai." The first basic extraction process extracts the SE clusters shown in the middle of the figure. This SE cluster clarifies that the words expressing this difference between the two English sentences are synonymous. Then, the second basic extraction procedure begins with a new situation in which the two SS groups are combined. The same process proceeds on the opposite language (Jpn) side .

As mentioned above, the iteration of the basic extraction procedure utilizing the previously extracted SE acquires an additional SE. This iteration continues until no additional SE is obtained.

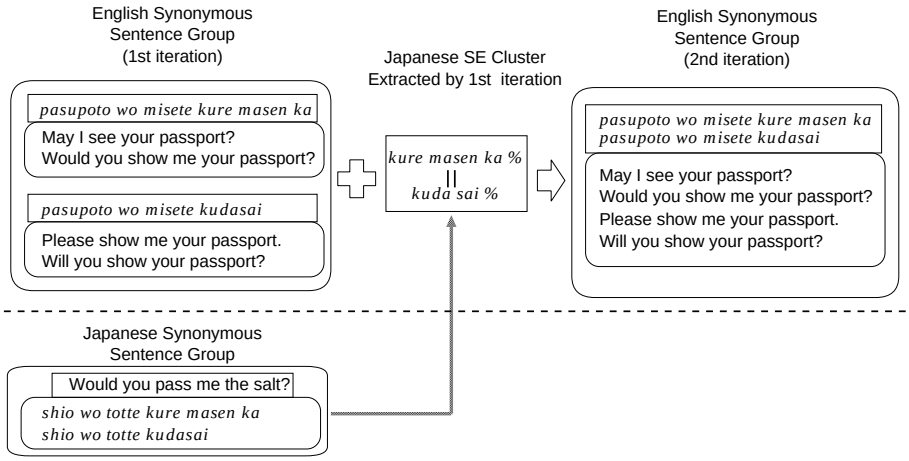

Figure 5.5. Integrating SS Group by Unification

3. Experiment

We next describe the experimental corpus in Section 3.1 and the extracted SE from the corpus in Section 3.2. We evaluated the effect of SE unification by comparing the performances of machine translations that use a corpus with or without unification. We describe the experimental result of applying unification to EBMT in Section 3.3 and to SMT in Section 3.4.

3.1 Corpus

The corpus we use is a collection of basic expressions in a travel situation (Takezawa et al., 2002). It is an English-Japanese sentence-aligned corpus.

We divide this corpus into learning data and test data. The learning data are used for both SE cluster extraction and experimental corpus-based machine translation. This means that the corpus used for machine translation simplifies its SE variety by itself. Table 5.1 shows the statistics of the learning data. The corpus contains many of the same sentences in each language. This corpus is favorable for our experiment because its sentences are relatively short and it has many identical sentences.

We have explained that our extraction method is based on synonymous “sentences.”

Table 5.1. Corpus for Synonymous Expression Extraction

	Japanese	English
# of Total Sentences	127,110	127,110
# of Different Sentences	89,615	85,839
Average # of Words in a Sentence	7.8	6.7

Table 5.2. Extracted Synonymous Expression Groups

		1st	10th	Expansion Ratio by Iteration (%)
SE Clusters	(Eng)	794	980	23.4%
Average Component Expressions	(Eng)	2.28	2.36	3.5%
SE Clusters	(Jpn)	1,110	1,251	12.7%
Average Component Expressions	(Jpn)	2.39	2.48	3.8%

However, this setting is too strict to apply to existing general parallel corpora, which have longer sentences and rarely have the same sentences in each language. When applying our method to these general parallel corpora, we have to use a finer unit, such as "phrase," rather than sentence. Techniques for acquiring phrase alignment from a parallel corpus have been proposed in many studies (Matsumoto et al., 1993; Watanabe et al., 2000).

3.2 Extracted Synonymous Expressions

We acquired synonymous expressions in English and Japanese by iterating the basic procedure. Figure 5.6 shows the number of extracted SE clusters and component expressions obtained by iteration. The number of SE clusters and expressions increases until the tenth iteration, and no change occurs after the eleventh iteration.

Table 5.2 shows the number of SE clusters and the average number of component expressions in an SE cluster by the first and the tenth iteration. Expansion ratios by

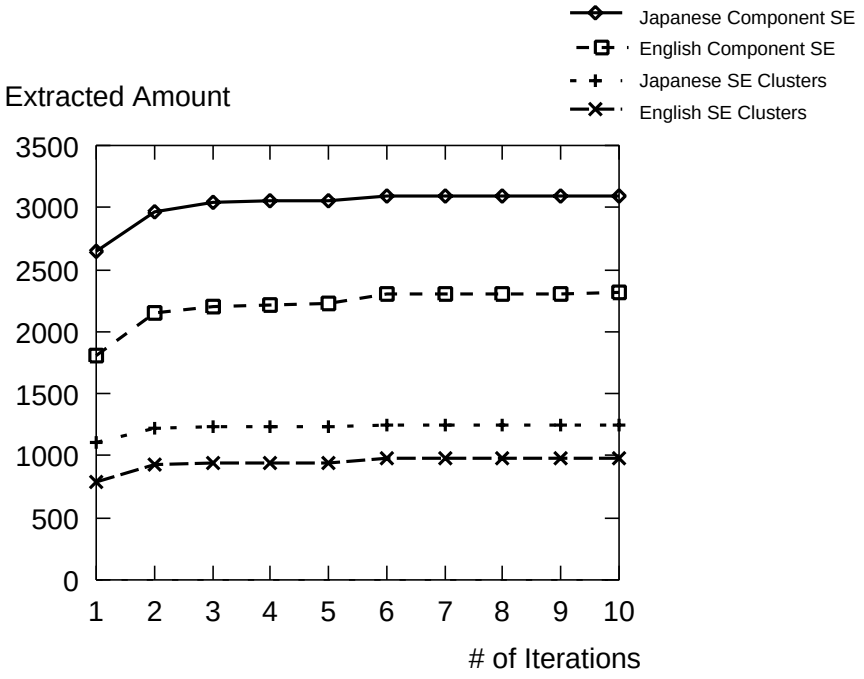

Figure 5.6. Extracted Synonymous Expressions by Iteration

iteration, and the ratio of the number in the 10th to that of the 1st, are also shown.

Iteration is effective for acquiring additional SE clusters rather than additional component expressions. From this corpus, more information is acquired by iteration for Japanese than for English in terms of both SE clusters and component expressions.

Examples of extracted SE clusters are shown in Figures 5.7 and 5.8. In the figures, the top expression in an SE cluster denotes a representative expression.

Our method extracts not only content word related SE but also function word related SE. Figure 5.9 shows the ratios of extracted expressions from major parts-of-speech (POS) [3]. The POS of function words are written in bold.

This table indicates the characteristics of the applied language. There are various expressions related to Auxiliary verb or Particle in Japanese. This reflects the phenomenon that there are various synonymous expressions to express the degree of politeness in Japanese conversation.

[3]Summation of ratios exceeds 100% since an expression involve several POS.

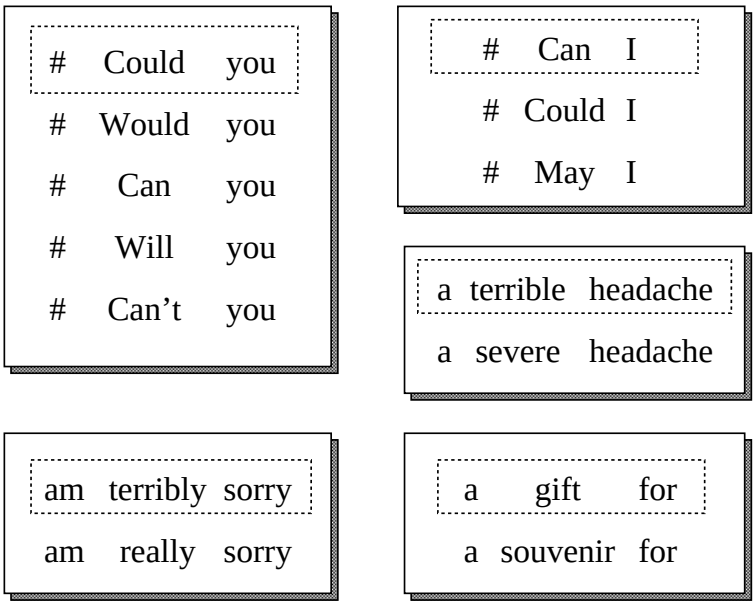

Figure 5.7. Example of English Extracted Synonymous Expression Groups

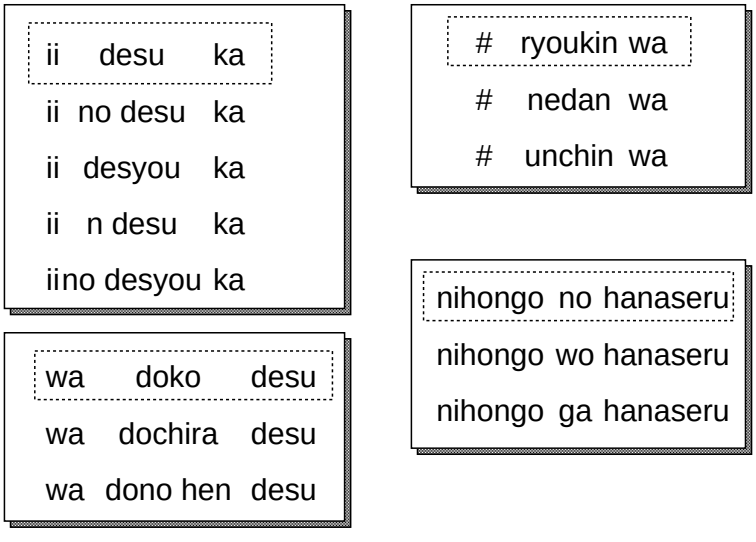

Figure 5.8. Example of Japanese Extracted Synonymous Expression Groups

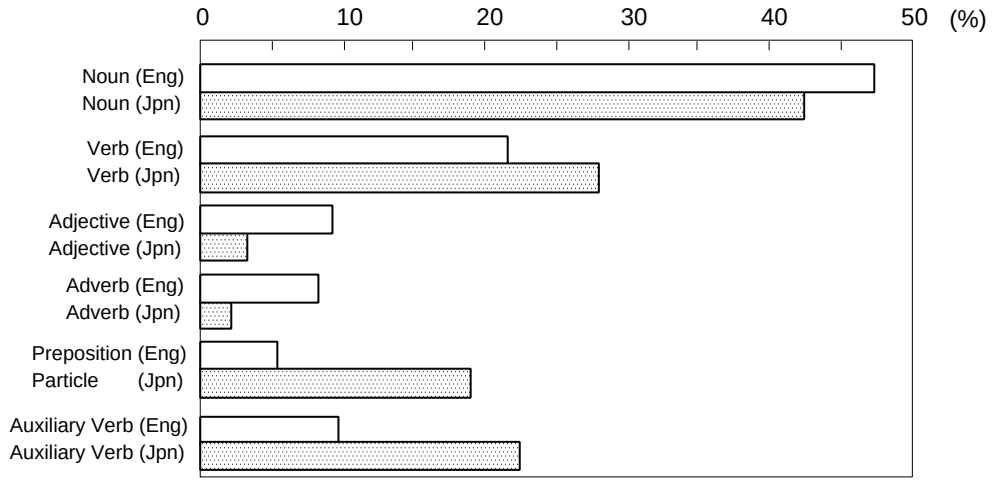

Figure 5.9. POS Distribution of Extracted Synonymous Expressions

We have compared our extracted SE clusters with Bunrui-Goi-Hyo[4]. There are 1,584 types of synonyms, and 976 (61.6%) of them are not contained in the thesaurus. Focusing on content words, 827 synonyms for content words are extracted, and 327 (37.5%) of them are not in the thesaurus. This demonstrates that our method extracts many synonymous expressions that are not covered by the thesaurus.

3.3 Application to EBMT

Experimental Environment

An overview of an experimental EBMT with a unification module is shown in Figure 5.10. SE clusters are extracted from an example corpus and installed in a similarity measuring module. We compared the performance of two settings: measuring similarity with (w/i) or without (w/o) unification. A similar sentence in the without-unification EBMT needs to be the same as the input sentence, while that in the with-unification EBMT needs to be the same after SE unification (Shimohata and Sumita, 2002). SE unification expands the coverage of translatable input sentences since it equates many literal expressions.

[4]Surrounding words of component expressions are stripped.

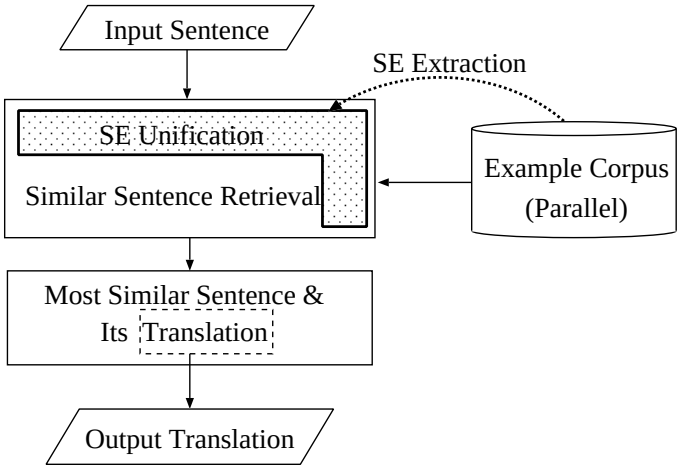

Figure 5.10. Experimental EBMT System

Table 5.3. Expansion of Translatable Inputs by Unifying Synonymous Expressions

Translation	# of Inp.	Translated Input Sentences		Exp. ratio (%)
		w/o Unif.	w/i Unif.	
Jpn-to-Eng	8,092	3,201	+280	8.7%
Eng-to-Jpn	7,564	2,890	+244	8.4%

Results

We evaluated the effect of SE unification in terms of the expansion of translatable sentences and the quality of additionally acquired translations. The expansion ratio of translatable sentences is determined from the ratio of additional translations by SE unification to the total translations without SE unification. The results are shown in Table 5.3. In Jpn-to-Eng translation, 3,201 of 8,092 input sentences were translated without unification. After unification, 280 of 4,891 untranslated sentences were additionally translated. Therefore, the expansion ratio is 8.7% (280/3201). Correspondingly, the expansion ratio in Eng-to-Jpn translation is 8.4%.

Each translation was evaluated by a human to determine whether it is appropriate.

Table 5.4. Evaluation of Translation Quality

		w/o Unif.	w/i Unif.
Jpn-to-Eng	Evaluated Sentences	500	280
	Proper Translations	486	267
	Precision (%)	97.2%	95.4%
Eng-to-Jpn	Evaluated Sentences	500	244
	Proper Translations	492	239
	Precision (%)	98.4%	98.0%

Table 5.4 shows a human-evaluation result for 500 translations randomly selected from among those acquired without SE unification and for all of the additional translations by SE unification. There is little degradation of translation quality in Eng-to-Jpn translation (0.4%) and Jpn-to-Eng translation (1.8%).

3.4 Application to SMT

We applied SE unification to target sentences in a learning corpus. This reduces the data sparseness problem in a translation model since it reduces the variety of synonymous expressions in the target language. We compared the performances of two SMT systems: using the learning corpus as it is (w/o unification) and using the learning corpus after SE unification (w/i unification) (Watanabe et al., 2002).

We used 240 Japanese sentences as input sentences and evaluated the output translation into four ranks: A (perfect), B (fair), C (acceptable), and D (nonsense). We considered sentences ranked as A, B, or C proper translations. Table 5.5 shows the translation quality of w/o unification and w/i unification systems. Precision increases 2.5%, which shows the effect of unification. Looking at the results in detail, the ratio of rank B increases while those of ranks A, C, and D decrease. The decrease of ranks C and D indicates that SE unification changes the translation model and improves performance. We believe the reason for the 1.3% decrease of rank A to be the fact that SE unification wipes out the delicate meaning of the original sentence. The removal of a delicate meaning is a side effect of SE unification.

Table 5.5. Difference in Translation Quality on SMT by Unifying Synonymous Expressions

	Rank			
	A	B	C	D
w/o Unification	29.2%	23.8%	17.1%	30.0%
w/i Unification	27.9%	28.8%	15.8%	27.5%

4. Related Work

In this section, we survey studies related to synonyms. A survey on thesauri, which are human-created synonym resources, is described in Section 4.1. A survey on automatic acquisition of resources is described in Section 4.2.

4.1 Thesauri

There are various human-created thesauri. The WordNet (Fellbaum, 1998) and Roget's Thesaurus (Roget, 1946) are well-known in English, and "Bunrui-Goi-Hyou" (The National Institute for Japanese Language, 1964) and "Nihongo-Goi-Taikei" (Ikehara et al., 1997) are well-known in Japanese.

A major problem in utilizing these existing thesauri for machine translation is that they have little information on function words. WordNet 2.0[5] contains information on 152,059 unique words, but the types of content are confined to nouns, verbs, adjectives, adverbs, and adjective satellites. Bunrui-Goi-Hyou contains information on about 32,600 words, but most of it refers to content words of nouns, verbs, and adjectives, and only 362 words refer to the other types. When measuring the similarity between two sentences, function word related expressions are as significant as content word related expressions.

In addition, human-created thesauri cover general information but rarely cover specific domains or specific usage. Our method can provide information on specific domains by extracting from a parallel corpus of the target domain. Another advantage of our

[5]Available from "http://www.cogsci.princeton.edu/~wn/wn2.0.shtml"

method is that it can extract SE from a translingual viewpoint.

4.2 Automatic Acquisition of Lexical Synonyms

There have been many studies on automatic acquisition of synonyms from a monolingual corpus. These studies overcome one of the shortcomings of human-created thesauri, i.e., covering synonyms in a specific domain. However, they still have the shortcoming of sparseness of function word related SE.

In these studies, the synonymy of two words is determined by the similarity of the contexts in which they apear. Similarity in the documents they are extracted from or the modification relation with other words is widely used to determine similarity (Manning and Schütze, 1999). (Lin, 1998) uses a 3-tuple consisting of a modifier, a modifiee, and the type of their relation. Extractable synonyms are confined to the words of nouns, verbs, and adjectives.

This approach is inappropriate for measuring the similarity of function word related synonyms. Since the role of function word is to add functional information to a modifying content word, the environment in which appears does not reflect its meaning.

(Barzilay and McKeown, 2001) proposed the automatic acquisition of paraphrases from a parallel corpus. The major difference from our method is that their extracted paraphrases do not contain contextual information. Though they state that paraphrases can improve the performance of multi-document summarization and sentence generation, the specific effect is not clarified.

5. Comparative Experiments on Various Language Pairs

This section describes the relation between a target language[6] and an intermediate language[7]. The relation is investigated through an experiment using English, Japanese, Korean, and Chinese text from BTEC. Among these, English and Japanese are used as target languages. The number of training sentences was 95,837. In this experiment,

[6] A language in which SE are extracted.

[7] A language used for binding target language sentences.

Original	Paraphrased	Evaluation
This is too big.	⇒ It is too big.	Same
Can we have lunch?	⇒ Can I have lunch	Same
How much is the set menu?	⇒ What's the set menu?	Diff.
I have an appointment.	⇒ I have a reservation.	Diff.
How much does it cost?	⇒ How much is it cost?	Inc.
Where are the toilets	⇒ Where is the toilets?	Inc.

Figure 5.11. Examples of Paraphrased Sentence Evaluation

iteration of base extraction could not be carried out since Korean and Chinese texts are not POS tagged texts but raw texts.

5.1 Evaluation Method

The synonymy of an extracted SE is evaluated by comparing paraphrased sentences with the meaning of the original sentence. Sentences in the test data are used as source sentences for paraphrasing. We call these sentences the original sentences. Sentences of the paraphrase language are paraphrased by the extracted SE. Paraphrasing is carried out by replacing non-standard expressions with standard expressions. All applicable SE are applied to sentences.

Evaluation is done by native speakers of the paraphrase languages. Evaluators are asked to label paraphrased sentences with any of the following marks.

Same Paraphrased sentence maintains the main meaning of the original sentence.

Diff Paraphrased sentence does not maintain the main meaning of the original sentence or is unnatural in itself.

Inc Paraphrased sentence is syntactically incorrect.

Figure 5.11 shows examples of synonymy evaluation. The ratio of "Same" evaluations to the all input sentences is taken as the precision of paraphrases (Prec.).

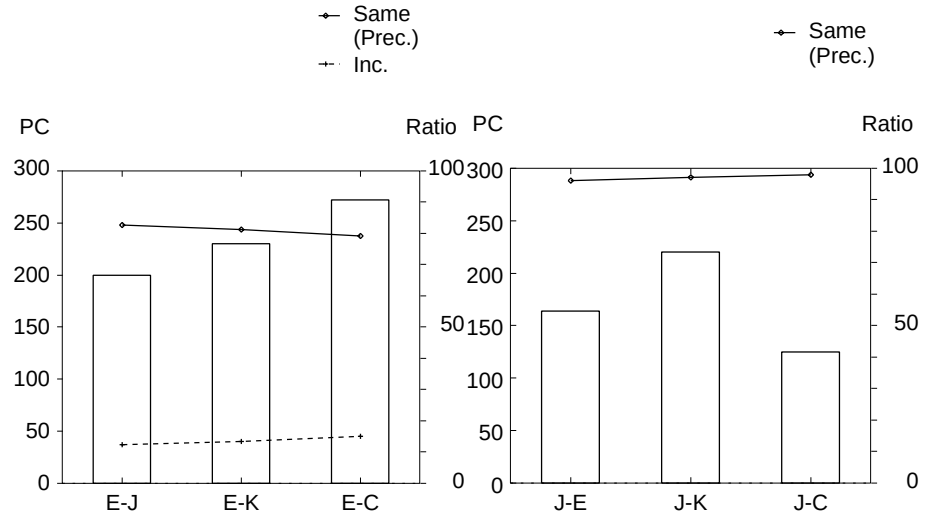

Figure 5.12. Results for Various Language Pairs

The number of SE is measured by the number of SE clusters. The applicability of paraphrasing is measured by the paraphrased ratio (PR), which is the ratio of paraphrased-sentences to all original sentences.

5.2 Results

Figure 5.12 shows the results of the experiment. In the figure, the histograms indicate the number of extracted SE, and the lines indicate the evaluations of extracted SE. The left side of the figure shows the results of English SE derived from Japanese, Korean, and Chinese. The right side of the figure shows the results of Japanese SE derived from English, Korean, and Chinese.

As for precision, it depends on the target language rather than the intermediate language. This suggests that the number of SE that reflect translingual viewpoints is relatively small.

As for the number of extracted paraphrases, both target and intermediate languages have a great influence. In general, the more target and intermediate languages are similar, the greater number of SE is acquired. Japanese SEs are more extensively extracted under J-K than under J-E and J-C. English SEs are more extensively extracted under

E-C than under E-J and E-K. Japanese and Korean are recognized as similar languages, but they are not similar to English and Chinese.

6. Conclusion

This chapter raised the issue of the variety of synonymous expressions and the problem this variety poses for natural language processing. We proposed a method for extracting synonymous expressions from a parallel corpus and for reducing their variety by unifying them.

Our SE has two advantages for application to machine translation:

1. They include expressions synonymous from the translingual viewpoint. This type of SE is effective for machine translation.

2. They include many function word related synonyms.

SE extraction is based on edit-operations between synonymous sentences. An expression consists of synonyms and the surrounding words. These surrounding words provide contextual information and improve SE extraction.

We demonstrated the effect of SE unification by applying it to two corpus-based machine translation methods: EBMT and SMT. SE clusters are extracted from a corpus used for two experimental MT systems. When applied to EBMT, they expand the coverage of translatable sentences by about 8.7% in J-to-E translation and 8.4% in E-to-J translation. The quality of additionally acquired translations degraded only slightly compared to translations acquired without our method. When applied to SMT, the extracted SEs improve translation accuracy by 2.5%.

Chapter 6

Retrieving a Similar Sentence from a Monolingual Corpus

Although machine translation (MT) technology has been undergoing development for several decades, its performance does not yet satisfy users' needs. Modifying an input sentence into a more translatable one, known as "pre-editing," is an important means of improving MT performance. (Bernth and Gdaniec, 2001) provided a guideline for the manual pre-editing of input sentences. (Mitamura and Nyberg, 2001) proposed a controlled language that is advantageous for MT. They also proposed a rewriting tool named KANTOO that supports an author in matching free input sentences to a controlled language. (Doi and Sumita, 2003) proposed an automatic pre-editing method that splits long input sentences. All of the previous works of pre-editing deal with partial modification of an input sentence.

We propose a novel pre-editing technique that incorporates similar sentence retrieval in MT to improve the translation of hard-to-translate[1] input sentences. The retrieval method has the advantage of relying only on a monolingual corpus, which is easy to prepare on large scale. Figure 6.1 shows an overview of our proposal. An input sentence can be classified as hard-to-translate or not by an MT system. If a given input sentence is hard to translate, the similar sentence retrieval function searches for the most similar sentence from a translatable sentence corpus[2] and provides it to the MT system. MT

[1] A "hard-to-translate" sentence refers to a sentence whose translation quality will probably be low when it is translated by an MT system.

[2] The corpus can be built by extracting translatable sentences from available monolingual corpora.

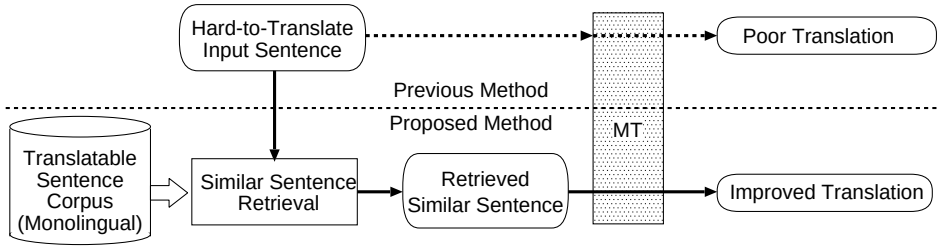

Figure 6.1. Improving Translation by Similar Sentence Retrieval

performance can be improved if the translation quality of retrieved sentences is better than that of the original sentences.

Our approach requires a translation quality measure to determine whether an input is hard-to-translate. Some MT systems can measure their translation quality by themselves. (Ueffing et al., 2003) proposed a method to estimate a confidence measure for statistical MT based on word graphs and N-best lists. The low-confidence translations correspond to hard-to-translate input sentences. Example-based MT systems can estimate their translation quality from the similarity distance between input and example sentences (Sumita, 2001). The parsing result of an input sentence is also useful.

The proposed method for measuring the similarity between input and candidate sentences[3] is based on research of the automatic evaluation of MT results. We adopt a metric based on the *common N-gram* after a comparative study of three methods: *common N-gram*, *common word sequence*, and *common word set*. (Section 1) Furthermore, we add two additional conditions to improve retrieval precision. (Section 2) We describe an experiment on applying similar sentence retrieval to a Japanese-to-English MT system in Section 3.

[3] "Candidate sentences" mean the sentences in a corpus to be retrieved.

1. Method for Measuring Similarity between Two Sentences

Our method for measuring similarity is based on research of automatic evaluation of MT results. Similarity measurement in automatic evaluation only depends on reference sentences[4] and does not require any other knowledge. In addition, many research attempts have demonstrated that automatic evaluation is strongly correlative with human evaluation.

Below, we give a brief overview of the research on automatic evaluation of MT and describe a comparative experiment among three methods and two additional conditions.

1.1 Overview of Automatic Evaluation of Machine Translation

In recent years, research on automatic evaluation of MT has become increasingly active. The basic idea of automatic evaluation is that a translation to be tested obtains higher score as it shares more common parts with several reference translations. There are three basic methods for measuring similarity between test translation and reference translations: common N-gram, word error rate (WER), and position-independent word error rate (PER).

The BLEU method (Papineni et al., 2002), which is one of the major methods, uses the common N-gram method. The similarity score is based on a common N-gram ratio of a test translation to the reference translations. The value of BLEU similarity ranges from 0 to 1. The higher the BLUE score is, the more similar a test sentence is. The method uses a brevity penalty to penalize short-length sentences. The NIST method (NIST, 2002), which is also widely used, is a revised version of the BLEU method.

The other two methods, WER and PER, are also often used (Tillmann et al., 1997). The WER is a length-normalized levenshtein distance. This metric is widely used to measure speech recognition errors. The PER is determined from the difference of the two word sets derived from the two sentences. This method differs from the WER method in that it ignores word order.

[4]They are a set of different proper translations of the test source sentence.

1.2 Basic Methods

Our method of measuring similarity is based on the ratios of common elements to the input and the candidate. If we regard the input sentence and the candidate sentence in our task as a single reference translation and a test translation, respectively, the ratio of common elements to the input sentence corresponds to precision, and that to the candidate sentence corresponds to recall. The definitions of precision, recall, and F-measure, which is the harmonic average of precision and recall, are given as follows.

$$\text{Precision} = \frac{\text{Common elements}}{\text{Elements in candidate}}$$

$$\text{Recall} = \frac{\text{Common elements}}{\text{Elements in input}}$$

$$\text{F-measure} = \frac{2\text{PR}}{\text{P} + \text{R}}$$

We compared the three basic methods, common N-gram, common word sequence, and common word set, which differ in their treatment of word order information. The common N-gram method takes local word order into account. The common word sequence takes word order through the sentence into account. The common word set method does not take word order information into account.

Common N-gram

The definition of the N-gram method is based on that of BLEU, which is a precision-based metric. Precision P is determined by the following equation.

$$P = \exp(\sum_{n=1}^{2} \frac{1}{n} \log(p_n))$$

p_n denotes precision for each n and can be determine by the following equation.

$$p_n = \frac{\sum_{N-gram' \in Candidate} Count_{clip}(N-gram')}{\sum_{N-gram \in Candidate} Count(N-gram)}$$

$Count(x)$ denotes the frequency of x in the candidate and $Count_{clip}(x)$ denotes the lower frequency of x in the input or the candidate.

Precision in the common N-gram method is also determined by the above. Recall in the common N-gram method is determined by replacing "$Candidate$" with"$Input$."

On the other hand, our method differs in the following points from the BLEU method.

1. To penalize short-length sentences, the BLEU method uses a brevity penalty while our method uses recall.

2. Our method uses only unigram and bigram, while the BLEU method uses unigram to 4-gram. This is because sentences in travel conversation have relatively short. The requirement in which a retrieved sentence needs to share any 4-gram is too strict for our target domain.

Common Word Sequence

This method is based on the longest common word sequences by using DP-matching (Cormen et al., 2001). The longest common word sequence means the word sequence whose component words appear in both sentences in the same order, but not necessarily consecutively.

This method has an inverse proportional relation with WER. Our preliminary experiment verified that the performances of the common word sequence metric and the WER metric are nearly equivalent.

Common Word Set

This method regards a sentence as a set of words, i.e. a bag-of-words, and defines a common element as a common word between input and candidate sentences. In other words, word order information is ignored. This method has an inverse proportional relation with PER.

1.3 Additional Conditions

We use two additional conditions to improve retrieval performance. The first condition, described in Section 1.3, rejects superficially resembling but not substitutive sentences. The second condition, described in Section 1.3, lessens the influence of style difference.

Excluding Sentences Having Additional Content Words

A preliminary experiment demonstrates that candidates that have additional content words from an input sentence are often dissimilar. This is because additional content words often work as an additional constraint on inputs and make candidates dissimilar

to the input. If the sentence "I'd like to reserve a table for two at **seven** tonight" is retrieved for the input "I'd like to reserve a table for two tonight," the retrieved sentence is dissimilar since the additional word "seven" causes significant misunderstanding. Therefore, an effective way to eliminate dissimilar sentences is to exclude sentences having additional content words from candidate sentences.

Content words are defined to include nouns, verbs, adjectives, adverbs, and numerals. Function words are defined to include particles, auxiliary verbs, and the copula. A compound word, as in the case of English "New York," "get off," and "two hundred dollars," is treated as a single word.

Reducing Function Word Weight

Input sentences have the characteristics of Japanese spoken language, while candidate sentences do not. Sentences of Japanese spoken language have a wide variety in expressions of function word. Expressions consisting of auxiliary verbs and particles have a wide variation according to the politeness level. Case particles are often omitted in Japanese spoken language, while not in other domains. These phenomena reduce the significance of function words.

In addition, although function words express important information such as case relation, modality, and tense, this information is often compensated by content words. For example, suppose that we have to guess a sentence having content words of "I," "leave," "wallet," and "taxi." We can guess that the sentence "I left my wallet in a taxi." is the most appropriate candidate, although we can imagine various other sentences.

These observations suggest that reducing the weight of function words is favorable in measuring sentence similarity. We verified this through a comparative experiment in which a function word weight is set to either 1.0 (equivalent to content words) or 0.4 (reduced). In the experiment, we fixed a content word weight as 1.0 and set a function word weight to 1.0 or 0.4. In an experimental candidate corpus, the average number of content words in a sentence was 3.0 words and that of function words was 4.3 words. The weight of 0.4 greatly reduced significance of function words. As for the common N-gram method, the weight of the N-gram goes to 0.4 when all component words are function words.

Input: The room next door is noisy. Please change my room.		
	Retrieved Sentence	Evaluation
1	Next door is too noisy. Could you get me a different room?	Similar
2	Could you change my room?	Similar
3	The room next door is noisy.	Dissimilar
4	Change please.	Dissimilar
5	Could you change my room tomorrow?	Dissimilar

Figure 6.2. Examples of Similarity Evaluation

1.4 Comparing Precision of Each Method and Additional Conditions

Setting

In this experiment, similar sentence retrieval modules received an input sentence and returned the sentence that has the highest F-measure among the candidate sentences. If more than one sentence has the same highest F-measure, one of them is selected randomly.

Definition of Similar Sentence Retrieval Precision

The performance of similar sentence retrieval is evaluated by "precision," as defined below.

$$\mathsf{Precision} = \frac{\#\ \mathsf{of\ Similar\ Retrieved\ Sentences}}{\#\ \mathsf{of\ Total\ Input\ Sentence}}$$

A "similar retrieved sentence" is defined as a substitutive sentence that allows a conversation to proceed. Evaluation examples are shown in Figure 6.2. In examples 3 and 4, the retrieved sentences are dissimilar due to missing significant information. In example 5, the retrieved sentence is dissimilar because its additional information leads to significant misunderstanding.

Table 6.1. Precision by Basic Methods and Additional Conditions

Excluding Additional	Reducing Weight	Precision by Basic Method(%)		
		N-gram	Word Sequence	Word Set
No	1.0	44.3%	43.4%	36.2%
Yes	1.0	53.0%	51.9%	51.3%
Yes	0.4	**54.9%**	53.8%	52.2%

Results

Precisions by combinations of each basic method and additional conditions are shown in Table 6.1. The conditions described in Sections 1.3 and 1.3 are referred to as "Excluding Additional" and "Reducing Weight" respectively. The highest precision of 54.9% was attained by using the N-gram method with two additional conditions.

A comparison of the first and second lines in Table 6.1 indicates a large effect by the condition of excluding additional. It improved precision by at least 8% for every basic methods. A comparison of the second and third lines indicates a small effect by the condition of reducing function word weight. Precision improved in every basic method but the improvement was no more than 2%. As for precision differences by the basic methods, these differences were small, at no more than 3% in the third result. We cannot determine the general superiority of any among the three methods from this experiment only. The same is true with an automatic MT evaluation research.

2. Filtering Retrieved Sentences

The retrieved sentence having the highest similarity score in the candidate corpus is not necessarily similar to the input sentence. We used two filtering conditions to exclude dissimilar retrieved sentences: number of missing content words and number of common content words.

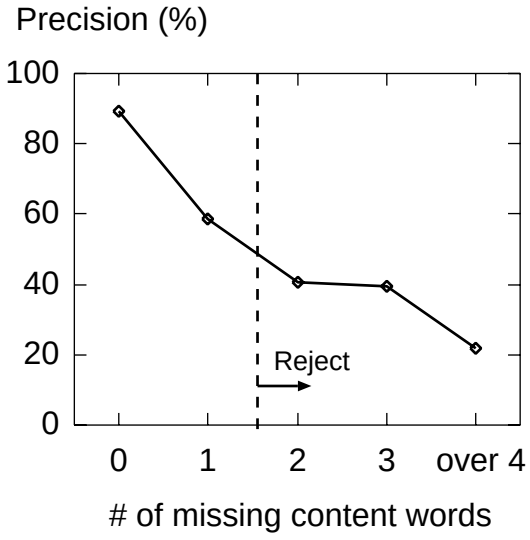

Figure 6.3. Precision by Number of Missing Content Words

2.1 Number of Missing Content Words

Utterances often contain redundant or easily guessable information. A retrieved sentence missing this information is still substitutive. This condition determines the maximum number of allowable missing content words.

Figure 6.3 shows precision by the number of missing content words. When all content words in an input sentence remained in a retrieved sentence, a high precision (89.1%) was attained. As number of missing content words increased, precision decreased. We defined that a retrieved sentence remains if it misses fewer than two content words. The precision with one missing content word was 58.8%.

2.2 Number of Common Content Words

We assume that a retrieved sentence sharing the main meaning with an input sentence is substitutive. This assumption suggests that if a retrieved sentence covers the main part of an input sentence, it is substitutive regardless of whether it misses other content words.

Figure 6.4 shows precision by the number of common content words. As the number

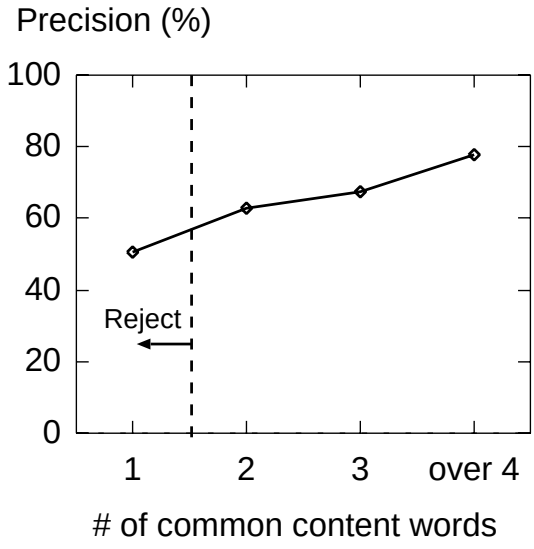

Figure 6.4. Accuracy by Number of Common Content Words

of common content words increased, precision increased. We defined that a retrieved sentence remains if it has more than one common content word. The precision when retrieved sentences had two content words was 63.0%, which is close to that defined in the previous section.

2.3 Results with Filtering

Figure 6.5 shows the results of similar sentence retrieval with filtering. Percentages in the figure indicate ratios to the items directly above. Retrieval ratio, the ratio of retrieved sentences to total inputs, was high (87.2%). Retrieval precision, the ratio of similar sentences to the retrieved sentences, was 60.4%, near the cut-off precisions defined in Sections 2.1 and 2.2.

This experiment also proved our hypothesis that function words are insignificant. We extracted the cases in which a retrieved sentence contains all content words but no function word of the input sentence. In this situation, it attains a high precision of 77.3%. This suggests that a coincidence of content words between an input and candidate sentence supports at least 77.3% precision regardless of function words.

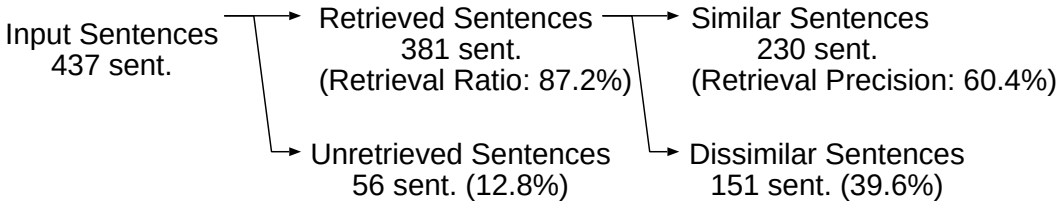

Figure 6.5. Retrieval Precision with Filtering

A detail analysis of dissimilar retrieved sentences clarifies that dissimilarity in many of them comes from lacking important information. To overcome this shortcoming, we have to distinguish important content words among all containing content words. Development of this distinguishment method is our future work.

3. Experiment on Application to Machine Translation

Figure 6.6 shows an overview of the experiment. Hard-to-translate sentences in the BTEC corpus are filtered out. As a result, 70,671 sentences remained, and they were used as the candidate corpus in this experiment. Then, 305 translatable sentences remained from 1,698 sentences in the travel conversation corpus. These sentences were provided to the experimental MT system as a collection of input sentences.

3.1 Experimental MT System

We used an example-based MT (EBMT) system in the experiment (Sumita, 2001). The basic idea of the EBMT system is that it retrieves sentences similar to input sentences from a parallel corpus and modifies the translation of the similar sentences to generate output translation. The similarity between the input sentence and example sentences is measured by edit distance. The weight of substitution is adjusted by the similarity of two words, which is based on the given thesaurus. Since translation quality derived from dissimilar sentences is low, the EBMT system outputs no translation if there is no similar example sentence in the corpus. Similar and dissimilar sentences are

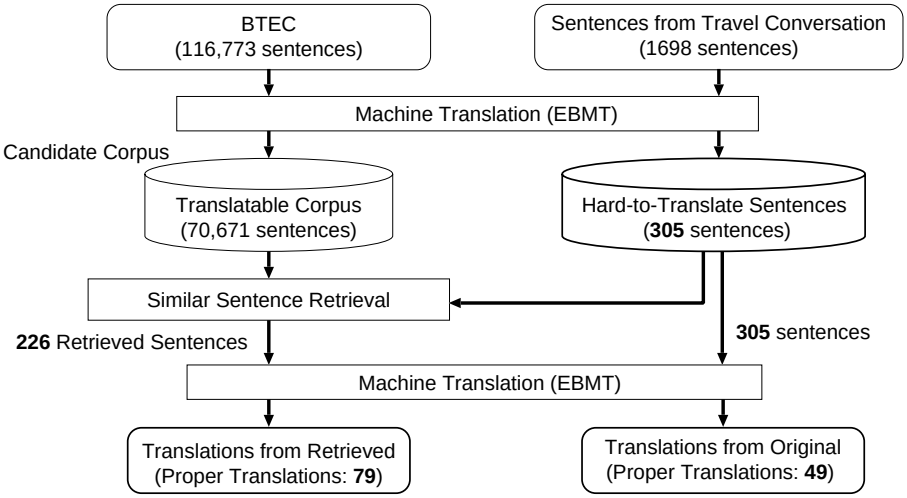

Figure 6.6. Overview of Experiment

distinguished by the predefined threshold of similarity distance.

Interestingly, the EBMT system also relies on similar sentence retrieval as with our proposed method. However, EBMT and our method differ in the types of corpus to be retrieved: EBMT deals with a parallel corpus and our method deals with a monolingual corpus. It is difficult to prepare a large-scale parallel corpus, and its construction is a great problem for EBMT. Our method enhances EBMT performance by utilizing a monolingual corpus, which is easy to build. Understandably, our method can be combined with any MT methods to improve their performance.

3.2 Results

The similar sentence retrieval module received 305 hard-to-translate sentences and returned 226 retrieved sentences (74.1% of the 305 sentences) as shown in Figure **??**.

We compared the similarity between input sentences and retrieved sentences manually. The evaluation criterion was the same as that described in Section 1.4. As a result, 99 of 226 retrieved sentences are proper. These sentences occupy 43.8% of the retrieved sentences and 32.5% of the total hard-to-translate sentences.

Then, we compared the translation quality derived from the original sentences and the

Table 6.2. Translation Quality of Original and Retrieved Sentences

Source Sentences	Proper Translations	Accuracy (%)
Retrieved	79	25.9%
Original	49	16.1%

retrieved sentences. The translations were evaluated as to whether they were proper or not. This evaluation criterion is the same as that in (Sumita, 2001) and is independent from that of the retrieved sentence evaluation. The results are shown in Table 6.2. In the table, accuracy denotes a ratio of proper translations to the 305 hard-to-translate input sentences. Input sentences that have no retrieved sentences are counted as improper translations. The result shows the accuracy of our method and the original to be 25.9% and 16.1% respectively, and our method attains an improvement of 9.8%.

4. Conclusions

We proposed a novel pre-editing technique of replacing a hard-to-translate input sentence with a similar translatable sentence. This strategy has the advantage of requiring only a monolingual corpus.

The development of similar sentence retrieval owes much to research on the automatic evaluation of MT. We adopted the common N-gram method among the three major methods after a comparative study. Furthermore, we added two conditions and filtering to our task. From an experiment applying the method to MT, the translation quality of hard-to-translate sentences was improved by 9.8%.

From the result described in Chapter 3, expansion ratio caused by sentential and phrasal paraphrasing is 5.23 (= 4.25 * 1.23). Our proposed method captures only a small part of this paraphrase variation. Expanding the coverage of paraphrases will be work that we do in the future.

Chapter 7

Rough Translation based on Similar Sentence Retrieval

EBMT is a promising translation method for speech-to-speech translation (S2ST) because of its robustness. However, there are two problems in applying EBMT to S2ST. One is that the translation accuracy drastically drops as input sentences become long. This is because as the length of a sentence becomes long, the number of retrieved similar sentences greatly decreases. This often results in no output when translating long sentences.

The other problem arises due to the differences in style between the input sentences and the example corpus. It is difficult to acquire a large volume of natural speech data since it requires much time and cost. Therefore, we cannot avoid using a corpus with pseudo speech-style text, which has a little different style from that of natural speech. This style difference makes retrieval of similar sentences difficult and degrades the performance of EBMT.

We propose example-based rough translation to overcome the above two problems of EBMT. Example-based rough translation is characterized by two points: (1) it allows missing unimportant information, and (2) it retrieves similar sentences based on content words and information of modality and tense. Tolerance of missing unimportant information brings robustness to the translation of long input sentences since this retrieval method substitutes similar short sentences for similar long sentences if there is no similar long sentence. Retrieval based on content word, modality, and tense brings

robustness to the style difference between the input sentences and the corpus. The style differences often appear in function words, and this retrieval strategy disregards almost all the information of function words except for the modality and tense information.

We describe the difficulties of applying EBMT to S2ST in Section 2. Then, we describe our purpose and retrieval method for meaning-equivalent sentences in Section 3 and a modification of the translation of meaning-equivalent sentences in Section 4. We report an experiment comparing our method with two other methods in Section 5. The experiment demonstrates the robustness of our method to the length of the input sentence and the style differences between the input sentences and the example corpus.

1. Related Work

The rough translation proposed in this paper is a type of EBMT (Sumita, 2001; Carl, 1999; Brown, 2000). The basic idea of EBMT is that sentences similar to the input sentences are retrieved from an example corpus and their translations become the basis of outputs. Here, let us consider the difference between our method and other EBMT methods by dividing similarity into a content-word part and a function-word part. In the content-word part, our method and other EBMT methods are almost the same. Content words are important information in a similarity measure process, and thesauri are utilized to extend lexical coverage. In the function-word part, our method is characterized by disregarding function words, while other EBMT methods still rely on them for the similarity measure. In our method, the lack of function word information is compensated by the semantically narrow variety in S2ST domains and the use of information on modality and tense. Consequently, our method gains robustness with regard to length and the style differences between the input sentence and the example corpus.

2. Difficulties of Applying EBMT to S2ST

2.1 Translation Degradation by Input Length

A major problem with machine translation, regardless of the translation method, is that performance drops rapidly as input sentences become longer. For EBMT, the longer

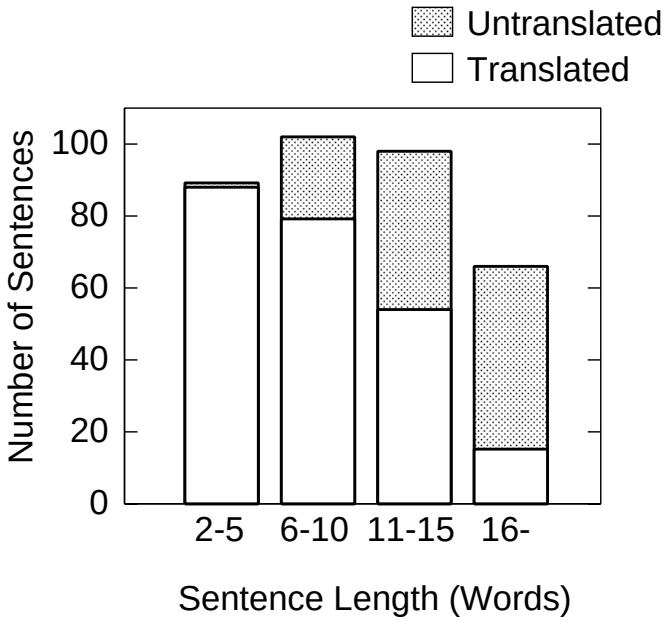

Figure 7.1. Distribution of Untranslated Input Sentences by Length

that input sentences become, the fewer similar example sentences exist in the example corpus. Figure 7.1 shows translation difficulty in long sentences in EBMT (Sumita, 2001). The EBMT system is given 591 test sentences and returns translation results as translated/untranslated. Untranslated means that no similar example sentence exists for the input sentence. Although this EBMT system was equipped with a large example corpus (about 170K sentences), it often failed to translate long input sentences.

3. Retrieving Meaning-equivalent Sentences for Rough Translation

In order to overcome the problems described in Section 2, we introduce an example-based rough translation strategy. Example-based rough translation has two key features: first, it uses a "meaning-equivalent sentence" which has a looser definition than the conventional "similar sentence" and second, it retrieves meaning-equivalent sen-

	Input Sentence	Unimportant?
1	Would you take a picture **of me**?	Yes
2	Would you take a picture **of this painting**?	No
3	Could you tell me a Chinese restaurant **around here**?	Yes
4	Could you tell me a **Chinese** restaurant around here?	No
5	My baggage was stolen from my room **while I was out**.	Yes
6	Please change my room **because the room next door is noisy**.	Yes

Figure 7.2. Examples of Unimportant Information

tences based on content words and information on modality and tense.

3.1 Meaning-equivalent Sentences

Meaning-equivalent sentences to an input sentence are defined as follows.

> A sentence that shares the main meaning with the input sentence despite missing some unimportant information. It does not contain information additional to that in the input sentence.

They bring robustness to the translation of long input sentences since sentences far shorter than input sentences can be retrieved as meaning-equivalent sentences. We assume that meaning-equivalent sentences (and their translations) are useful enough for S2ST, since unimportant information rarely disturbs the progress of dialogs and can be recovered in the following dialog if needed.

Important information is subjectively recognized mainly due to one of two reasons: (1) It can be guessed from the general situation, or (2) It does not add significant information to the main meaning.

Figure 7.2 shows examples of unimportant/important information. The information to be examined is written in bold. The information "*of me*" in (1) and "*around here*" in (3) can be guessed from the general situation, while the information "*of this painting*" in (2) and "*Chinese*" in (4) would not be guessed since it denotes a special object. The

subordinate sentences in (5) and (6) are regarded as unimportant since they have small significance and are omittable.

3.2 Basic Idea of Retrieval of Meaning-equivalent Sentences

The retrieval of meaning-equivalent sentences depends on content words and basically does not depend on function words. Independence from function words brings robustness to the difference in styles.

However, function words include important information for sentence meaning: the case relation of content words, modality, and tense. Lack of case relation information is compensated by the nature of the restricted domain. A restricted domain, as a domain of S2ST, has a relatively small lexicon and meaning variety. Therefore, if content words included in an input sentence are given, their relation is almost always determined in the domain. Modality and tense information is extracted from function words and utilized in classifying the meaning of a sentence (described in Section 3.3).

This retrieval method is similar to information retrieval in that content words are used as clues for retrieval (Frakes and Baeza-Yates, 1992). However, our task has two difficulties: (1) Retrieval is carried out not by documents but by single sentences. This reduces the effectiveness of word frequencies. (2) The differences in modality and tense in sentences have to be considered since they play an important role in determining a sentence's communicative meaning.

3.3 Features for Retrieval

Content Words

Words categorized as either noun[1], adjective, adverb, or verb are recognized as content words. Interrogatives are also included. Words such as particles, auxiliary verbs, conjunctions, and interjections are recognized as function words.

We utilize a thesaurus to expand the coverage of the example corpus. We call the relation of two words that are the same "identical" and words that are synonymous in the given thesaurus "synonymous."

[1]Number and pronoun are included.

Table 7.1. Clues for Discriminating Modalities in Japanese

Modality	Clues
Request	*tekudasai* (auxiliary verb) *teitadakeru* (auxiliary verb)
Desire	*shi-tai* (expression) *te-hoshii* (expression) *negau* (verb)
Question	*ka* (final particle) *ne* (final particle)
Negation	*nai* (auxiliary verb or adjective) *masen* (auxiliary verb)

Tense	Clues
Past	*ta* (auxiliary verb)

Modality and Tense

The meaning of a sentence is discriminated by its modality and tense, since these factors obviously determine meaning. We defined two modality groups and one tense group by examining our corpus. The modality groups are ("request", "desire", "question", "confirmation", "others") and ("negation","others"). The tense group is ("past", "others"). These modalities and tenses are distinguished by surface clues, mainly by particles and auxiliary verbs. These distinguishing rules were manually developed in several weeks. Table 7.1 shows some of the clues used for discriminating modalities in Japanese. Sentences having no clues are classified as "others". Figure 7.3 shows sample sentences and their modality and tense. Clues are underlined.

[2]Japanese content words are written in sans serif style and Japanese function words in *italic* style. Space characters are inserted into word boundaries in Japanese texts.

[3]The value "others" in all modality/tense groups is omitted.

Sentence[2]	Modality	Tense[3]
hoteru *o* yoyaku *shi* *tekudasai* (Will you reserve this hotel?)	request	—
hoteru *o* yoyaku *shi tai* (I want to reserve this hotel.)	desire	—
hoteru *o* yoyaku *shi mashi ta ka*? (Did you reserve this hotel?)	question	past
hoteru *o* yoyaku *shi tei masen* (I do not reserve this hotel.)	negation	—

Figure 7.3. Sentences and Their Modality and Tense

A sentence that satisfies the conditions below is recognized as a meaning-equivalent sentence.

3.4 Retrieval and Ranking

1. It has the same modality and tense as the input sentence.
2. All content words are included (identical or synonymous) in the input sentence. This means that the set of content words of a meaning-equivalent sentence is a subset of the input sentence.
3. At least one content word is included (identical) in the input sentence.

If more than one sentence is retrieved, we must rank them to select the most similar one. We introduce "focus area" in the ranking process to select sentences that are meaning-equivalent to the main sentence in complex sentences. We set the focus area as the last N words from the word list of an input sentence. N denotes the number of content words in meaning-equivalent sentences. This is because main sentences in complex sentences tend to be placed at the end in Japanese.

The retrieved sentences are ranked by the conditions described below. Conditions are described in order of priority. If there is more than one sentence having the highest

score under these conditions, the most similar sentence is selected randomly.

C1: # of identical words in focus area.
C2: # of synonymous words in focus area.
C3: # of identical words in non-focus area.
C4: # of synonymous words in non-focus area.
C5: # of common function words.
C6: # of different function words.
(the fewer, the higher priority)

Figure 7.4 shows an example of conditions for ranking. Content words in a focus area of the input sentence are underlined and function words are written in italic.

4. Modification

The sentence with the highest score among the retrieved meaning-equivalent sentences and its translation are taken. If the retrieved sentence has a synonymous word with the input sentence, the synonymous word in the translation of the retrieved sentence is replaced by the translation of the corresponding word in the input sentence.

Figure 7.5 shows the replacement of synonymous words in the translation of the retrieved sentence. The sentence "baggu *o* nusuma *re mashi ta*" is retrieved as the most meaning-equivalent sentence of the input "toranku *ga* nusuma *re tan desu*." The word "baggu"(bag) in the retrieved sentence and the word "toranku"(trunk) in the input are synonymous. Therefore, the translation of the retrieved sentence "My bag was stolen" is modified by replacing the word "bag" with "trunk," and the modified translation becomes the output. In this process, the word alignment between the meaning-equivalent sentence and its translation is automatically determined based on a translation dictionary.

[5]Words are converted to base form.

Input
gaishutsu *shi teiru* aida *ni*,
(While I was out),
kaban *o* nusuma *re mashi ta*
(my baggage was stolen.)

Meaning-equivalent Sentence
baggu *o* nusuma *re ta*
(My bag was stolen).

C1	nusumu[5]	1
C2	(kaban = baggu)	1
C3	-	0
C4	-	0
C5	*o, re, ta*	3
C6	*suru, teiru, ni, masu*	4

Figure 7.4. Example of Conditions for Ranking

5. Experiment

5.1 Test Data

The BTEC is divided into example data (Example) and test data (Concise) by extracting test data randomly from the whole set of data. The later part was used for an experiment applying similar sentence retrieval to MT (Section 5).

In addition to this, we used the TDC for another set of test data (Takezawa, 1999). This corpus contains dialogs between a traveler and a hotel receptionist. It is used to test the robustness against styles. We call this test corpus "Conversational."

We use sentences including more than one content word among the three corpora. The statistics of the three corpora are shown in Table 7.2.

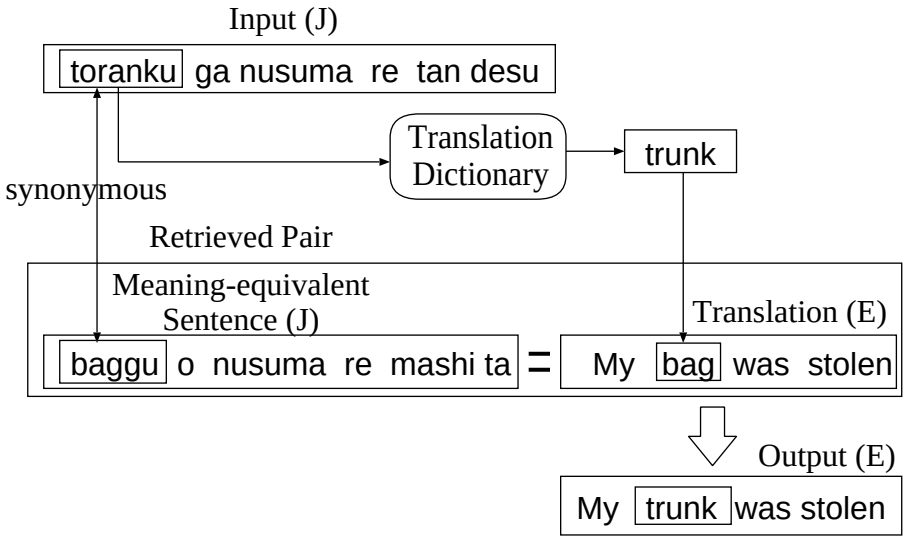

Figure 7.5. Replacement of Synonymous Words

The thesaurus used in the experiment was "Kadokawa-Ruigo-Jisho" (Ohno and Hamanishi, 1984). Each word has a semantic code consisting of three digits, that is, this thesaurus has three hierarchies. We defined "synonymous" words as sharing exact semantic codes.

Table 7.2. Statistics of the Corpora

Corpus	# of Sentences	Average Length
Example	92,397	7.4
Concise	1,588	6.6
Conversational	800	10.1

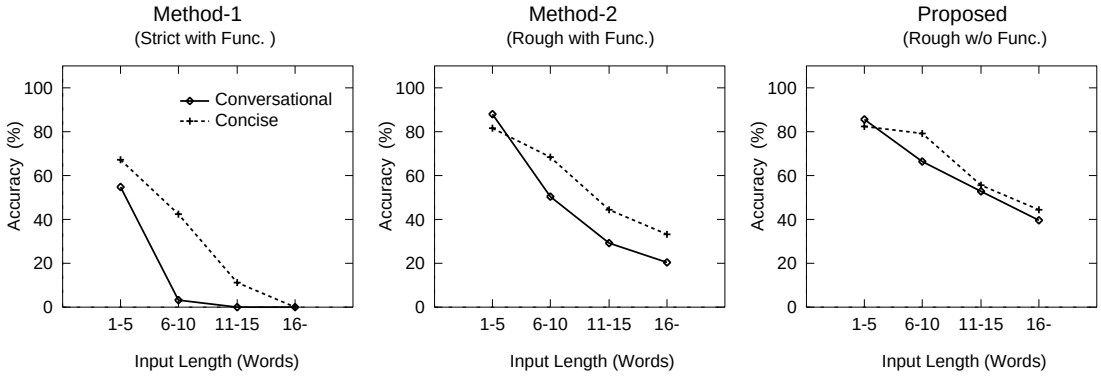

Figure 7.6. Retrieval Accuracy

5.2 Compared Methods for Meaning-equivalent Sentence Retrieval

We use two retrieval methods to show the characteristic of the proposed method. The first method (Method-1) adopts "strict" retrieval, which does not allow missing words in input. The method takes function words into account on retrieval. This method corresponds to the conventional EBMT method. The second method (Method-2) adopts "rough" retrieval, which does allow missing words in input, but still takes function words into account. The translation process in these two methods and proposed method is the same.

5.3 Accuracy of Meaning-equivalent Sentence Retrieval

Evaluation was carried out by judging whether the retrieved sentences are meaning-equivalent to the input sentences. The sentences were marked manually as meaning-equivalent or not by a Japanese native-speaker. Figure 7.6 shows the retrieval accuracy of the three methods with the concise and conversational style data. Retrieval accuracy is defined as the ratio of the number of correctly equivalent sentences to that of the total input sentences. The input sentences are classified into four types by their word length.

The performance of Method-1 reflects the narrow coverage and style-dependency of

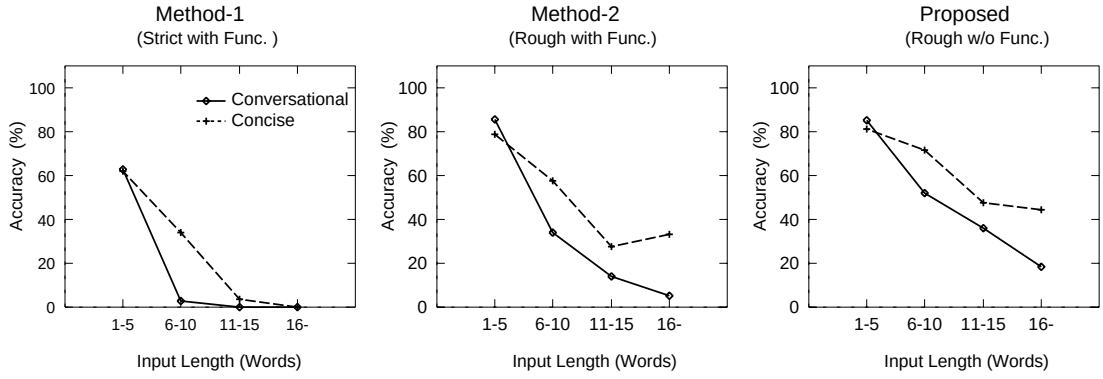

Figure 7.7. Translation Accuracy

conventional EBMT. The longer that the input sentences become, the more steeply its performance degrades in both styles. The method can retrieve no similar sentence for input sentences longer than eleven words in conversational style.

Method-2 adopts a "rough" strategy in retrieval. It attains higher accuracy than Method-1, especially with longer input sentences. This indicates the robustness of the rough retrieval strategy to longer input sentences. However, the method still has an accuracy difference of about 15% between the two styles.

The accuracy of the proposed method is better than that of Method-2, especially in conversational style. The accuracy difference in longer input sentences becomes smaller (about 4%) than that of Method-2. This indicates the robustness of the proposed method to the differences between the two styles.

5.4 Translation Accuracy

Translation accuracy was judged by an English native-speaker. It is defined as the ratio of the number of roughly appropriate translations to that of the total input sentences. Roughly appropriate translations correspond to translations of meaning-equivalent sentences. Figure 7.7 shows the translation accuracy of the three methods with the concise and conversational style data by input length. As done with retrieval accuracy, the translation accuracy from the proposed method was improved in both long input sentences and conversational styles.

Table 7.3. Overall Accuracy with Conversational Data

Method	Retrieval Accuracy (%)	Translation Accuracy (%)
Method-1	25.3%	24.2%
Method-2	54.2%	42.5%
Proposed	63.6%	50.7%

Table 7.3 shows the overall accuracy for all sentences with conversational data for retrieval accuracy and translation accuracy. The accuracy drop between the retrieval and translation of the rough methods (Method-2 and Proposed) is much larger than that of the strict method (Method-1). One reason for this larger drop is that a context discrepancy between the input sentence and the translation of a meaning-equivalent sentence occurs in the rough methods. This is because unimportant information, which is ignored in rough retrieval methods, has the effect of avoiding the retrieval of sentences having a different context from that of the input sentence. However, retrieval relying on unimportant information degrades total translation accuracy as shown in Table 7.3. In order to reduce the translation accuracy drop in the rough methods, it is effective to introduce contextual information, such as the scene of the utterance and the type of speaker, in the retrieval process (Yamada et al., 2000).

6. Conclusion

We proposed example-based rough translation for S2ST. It aims not at exact translation with narrow coverage but at rough translation with wide coverage. For S2ST, we assume that this translation strategy is sufficiently useful.

Rough translation is based on meaning-equivalent sentences that have the same main meaning as the input sentence despite missing some unimportant information. The retrieval of meaning-equivalent sentences is based on content words, modality, and tense. This strategy of rough translation brings robustness to the input length and the style differences between input sentences and the example corpus. An experiment on

travel conversation demonstrated these advantages.

Most MT systems aim to achieve exact translation, but unfortunately they often output bad or no translation for long conversational speeches. Rough translation achieves robustness in translating such input sentences. This method compensates for the shortcomings of conventional MT and makes S2ST technology more practical.

Chapter 8

Conclusion

Paraphrases, which express the meaning by using different words, play an important role in human communication since they bring a rich expressiveness to natural language. However, they also complicate a language and cause difficulty for natural language processing (NLP) systems.

In this thesis, we described a method for acquiring paraphrases and utilizing them to improve MT performance. The proposed methods are corpus-based approaches that acquire function word related expressions. We focused on lexical and sentential paraphrasing, which proved dominant in our investigation. Our method improved MT performance in both lexical paraphrasing and sentential paraphrasing.

1. Summary

- Analysis of Human Paraphrasing
  The author proposed paraphrase classification according to paraphrasing range: sentential, phrasal, or lexical. Two human paraphrasers considered sentences derived from travel conversations and made paraphrased sentences by using the three paraphrasing types. Analysis of paraphrased sentences indicated that the expansion ratios were in the order of lexical (11.15), sentential (4.25), and phrasal (1.23). An English sentence can be paraphrased into 60.36 other sentences on average.

- Extracting Lexical Paraphrases from a Parallel Corpus

Our proposed method is advantageous for MT since it extracts function word related paraphrases and paraphrases that are synonymous from a translingual viewpoint. Extracted paraphrases simplify texts in a corpus by replacing synonymous expressions with a single expression.

An experiment proved that extracted paraphrases improve the performances of two different corpus-based MT systems. Coverage of translatable sentences in EBMT expanded 8.7% in J-to-E translations and 8.4% in E-to-J translations. The translation quality of SMT improved 2.5% with our method.

- Retrieving Similar Sentences from a Monolingual Corpus
  A method for similar sentence retrieval, that is, sentential paraphrasing, was proposed. This method has an advantage: it only needs a monolingual corpus, which is easy to prepare. Comparative studies indicate that the adopted N-gram co-occurrence metrics is better for measuring similarity than metrics based on either longest common word sequence or common word set. However, the performance gap between the three metrics is small. We used two additional conditions that excluded sentences having additional content words and reduced function word weight. These two conditions had a greater effect on retrieval performance.

  When an original input sentence cannot be translated by an MT system, the sentence is replaced with a similar retrieved sentence. An experiment demonstrated that our proposed method improves EBMT performance.

- Effect of Human Paraphrasing on MT
  We proposed three paraphrasing methods: concise paraphrasing for short sentences and summary and segment paraphrasing for long sentences. Sentences in spoken language are paraphrased according to the three paraphrasing methods. Then we compared translation quality derived from original sentences and from paraphrased sentences. The experimental results indicate that paraphrasing is effective for EBMT systems while it has little effect on SMT systems. The detailed analysis finds that negative paraphrasing occurs frequently, which cancels the effect of positive paraphrasing. This result suggests that we should adapt paraphrasing rules for applying MT in order to gain steady improvement.

2. Future Work

Paraphrasing is a very attractive research topic that is deeply concerned with human language, and its research achievements will produce great benefits. We investigated human paraphrasing, proposed two methods to acquire paraphrases, and utilized them to improve MT performance. However, these achievements only exploited a minor aspect of paraphrasing. In the future we plan to continue to exploit paraphrases by extending our research into other applications and evaluation methodology.

- Other Applications
  Paraphrasing technology can provide great benefits for various purposes. The proposed paraphrasing methods were applied to MT and showed their effects. We plan to apply our technique to such NLP applications as information retrieval and automatic summarization. Furthermore, we intend to use our methods for such human text processing as paraphrasing to improve readability.

- Automatic Evaluation of Paraphrases
  Presently, paraphrased sentences are evaluated by humans. However, human evaluation is time-consuming and creates a bottleneck for development. Automatic evaluation of paraphrasing is necessary for rapid development. Fortunately, research into automatic MT evaluation and automatic summarization has advanced in recent years. We believe that the achievements of automatic MT evaluation will lead to automatic paraphrase evaluation.

Bibliography

R. Barzilay and K. R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proc. of the 39th Association for Computational Linguistics (ACL)*, pages 50–57.

Arendse Bernth and Claudia Gdaniec. 2001. MTranslatability. *Machine Translation*, 16(3):175–218.

P. F. Brown, J. Cocke, S. D. Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

R. D. Brown. 2000. Automated generalization of translation examples. In *Proc. of the 18th International Conference on Computational Linguistics (COLING-2000)*, pages 125–131.

M. Carl. 1999. Inducing translation templates for example-based machine translation. In *Proc. of the 7th Machine Translation Summit (MT Summit VII)*, pages 250–258.

J. Carroll, G. Minnen, D. Pearce, Y. Canning, S. Devlin, and J. Tait. 1999. Simplifying text for language-impaired readers. In *Proc. of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 269–270.

T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. 2001. *Introduction to Algorithms*. MIT Press.

T. Doi and E. Sumita. 2003. Input sentence splitting and translating. *Proc. of Workshop on Building and Using Parallel Texts, HLT-NAACL 2003*, pages 104–110.

F. Duclaye, F. Yvon, and O. Collin. 2003. Learning paraphrases to improve a question-answering system. In *Proc. of the EACL workshop on Natural Language Processing for Question-Answering*, pages 35–41.

C Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

A. Finch, Y. Akiba, and E. Sumita. 2004. Using a paraphraser to improve machine translation evaluation. In *Proc. of the 1st International Joint Conference on Natural Language Processing (IJCNLP2004)*.

W. B. Frakes and R. Baeza-Yates, editors. 1992. *Information Retrieval Data Structures & Algorithms*. Prentice Hall.

A. Fujita and K. Inui. 2001. Paraphrasing of common nouns to their synonyms using definition statements. In *Proc. of the 7th Annual Meeting of the Association for Natural Language Processing*, pages 331–334.

A. Fujita and K. Inui. 2003. Exploring transfer errors in lexical and structural paraphrasing. *Information Processing Society of Japan (IPSJ) Journal*, 44(11):2826–2838. (in Japanese).

S. Ide, T. Ogino, A. Kawasaki, and S. Ikuta. 1986. *Nihonjin to Amerikajin no Keigo Koudo*. Nan'un-Do Publishing. (in Japanese).

S. Ikehara, M. Miyazaki, S. Shirai, A. Yokoo, H. Nakaiwa, K. Ogura, Y. Oyama, and Y. Hayashi, editors. 1997. *Nihongo Goi Taikei*. Iwanami Shoten. (in Japanese).

K. Imamura, Y. Akiba, and E. Sumita. 2001. Paraphrasing of Japanese translation set using hierarchical phrase alignment. In *Proc. of Workshop Program of the 7th Annual Meeting of the Association for Natural Language Processing*, pages 15–20. (in Japanese).

K. Imamura, E. Sumita, and Y. Matsumoto. 2003. Feedback cleaning of machine translation rules using automatic evaluation. In *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-03)*, pages 447–454.

K. Inui, A. Fujita, T. Takahashi, R. Iida, and T. Iwakura. 2003. Text simplification for reading assistance: A project note. In *Proc. of 2nd International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP2003)*, pages 9–16.

C. Jacquemin, J. Klavans, and E. Tzoukermann. 1997. Expansion of multi-word terms for indexing and retrieval using morphology and syntax. In *Proc. of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97)*.

D. Jurafsky and J. H. Martin, editors. 2000. *Speech and Language Processing*. Prentice Hall.

N. Kaji, D. Kawahara, S. Kurohashi, and S. Sato. 2002a. Learning verb paraphrasing rules from a dictionary and a corpus. In *Proc. of the 8th Annual Meeting of the Association for Natural Language Processing*, pages 331–334. (in Japanese).

N. Kaji, D. Kawahara, S. Kurohashi, and S. Sato. 2002b. Verb paraphrase based on case frame alignment. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pages 215–222.

G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto. 2003. Creating corpora for speech-to-speech translation. In *Eurospeech-2003*, pages 381–384.

Y. Kim and T. Ehara. 1994. An automatic sentence breaking and subject supplement method for J/E machine translation. *Information Processing Society of Japan (IPSJ) Journal*, 35(6):1018–1028. (in Japanese).

Y. Kinjo, K. Aono, K. Yasuda, T. Takezawa, and G. Kikui. 2003. Correction of Japanese paraphrase data on basic expression of travel conversation. In *Proc. of the 9th Annual Meeting of the Association for Natural Language Processing*, pages 101–104. (in Japanese).

K. Kondo and M. Okumura. 1997. Summarization with dictionary-based paraphrasing. In *Proc. of the 4th Natural Language Processing Pacific Rim Symposium (NLPRS)*, pages 649–652.

G. Kowalski. 1997. *Information Retrieval Systems: Theory and Implementation*. Kluwer Academic Publishers.

G. Lazzari. 2002. The V1 framework program in Europe: Some thoughts about speech to speech translation research. In *Proc. of 40th ACL Workshop on Speech-to-Speech Translation*, pages 129–135.

D. Lin. 1998. Automatic retrieval and clustering of similar words. In *Proc. of COLING-ACL 98*, pages 768–774.

C. D. Manning and H. Schütze, editors, 1999. *Foundations of Statistical Natural Language Processing*, chapter Lexical Acquisition, pages 265–314. MIT Press.

S. Marumoto, T. Shirado, and H. Isahara. 2003. Statistical analysis on impression of honorific expressions. In *Proc. of the 9th Annual Meeting of the Association for Natural Language Processing*, pages 565–568. (in Japanese).

Y. Matsumoto, H. Ishimoto, and T. Utsuro. 1993. Structural matching of parallel texts. In *the 31st Annual Meeting of the ACL*, pages 23–30.

F. Metze, J. McDonough, H. Soltau, C. Langley, A. Lavie, L. Levin, T. Schultz, A. Waible, R. Cattoni, G. Lazzari, N. Mana, F. Pianesi, and E. Pianta. 2002. The nespole! speech-to-speech translation system. In *Proc. of Human language technology (HLT)*.

T. Mitamura and E. Nyberg. 2001. Automatic rewriteng for controlled language translation. *Proc. of NLPRS-2001 Workshop on Automatic Paraphrasing: Theories and Applications*, pages 1–12.

M. Nagao. 1981. A framework of a mechanical translation between Japanese and English by analogy principle. In *Artificial and Human Intelligence*, pages 173–180.

T. Nakao, J. Hibiya, and N. Hattori. 1997. *Shakai Gengogaku Gairon*. Kuroshio Publishing.

NIST. 2002. http://nist.gov/speech/tests/mt/.

S. Ohno and M. Hamanishi, editors. 1984. *Ruigo-Shin-Jiten*. Kadokawa. (in Japanese).

K. Ohtake and K. Yamamoto. 2001. Paraphrasing honorifics. In *Proc. of Automatic Paraphrasing: Theories and Applications (NLPRS2001 Workshop)*, pages 13–20.

K. Ohtake and K. Yamamoto. 2003. Applicability analysis of corpus-derived paraphrases toward example-based paraphrasing. In *Proc. of the 17th Pacific Asia Conference on Language, Information and Computation*, pages 380–391.

B. Pang, K. Knight, and D. Marcu. 2003. Syntax-based alignment of multiple translations: existing paraphrases and generating new sentences. In *Proc. of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 102–109.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pages 311–318.

P.M. Roget. 1946. *Roget's International Thesaurus*. Thomas Y. Crowell.

M. Shimohata and E. Sumita. 2002. Identifying expressions from a bilingual corpus for example-based machine translation. In *Proceedings of the workshop on machine translation in Asia*, pages 20–25.

M. Shimohata, T. Takezawa, and G. Kikui. 2003a. Construction of english paraphrase corpus on travel conversation and its analysis (in Japanese). In *Proc. of the Information Technology Letters*, pages 83–85.

M. Shimohata, T. Watanabe, E. Sumita, and Y. Matsumoto. 2003b. Extracting synonymous expressions from a parallel corpus for machine translation (in Japanese). *Information Processing Society of Japan (IPSJ) Journal*, 44(11):2854–2863.

M. Shimohata, E. Sumita, and Y. Matsumoto. 2004a. Building a paraphrase corpus for speech translation. In *Proc. of 4th international conference on language resources and evaluation (LREC)*, pages 453–457.

M. Shimohata, E. Sumita, and Y. Matsumoto. 2004b. A method for retrieving a similar sentence and its application to speech translation. *Journal of Natural Language Processing (to appear)*. (in Japanese).

Y. Shinyama, S. Sekine, K. Sudo, and R. Grishman. 2002. Automatic paraphrase acquisition from news articles. In *In Proc. of the 2nd International Conference on Human Language Technology Research*.

T. Shirado, S. Marumoto, and H. Isahara. 2003. Quantitative analyses of misusage of polite expressions. *Mathematical Linguistics*, 24(2):65–80. (in Japanese).

S. Shirai, S. Ikehara, T. Kawaoka, and Y. Nakamura. 1995. Effects of automatic rewriting of the source language within a Japanese to English MT system. *Information Processing Society of Japan (IPSJ) Journal*, 36(1):12–21. (in Japanese).

F. Sugaya, T. Takezawa, G. Kikui, and S. Yamamoto. 2002. Proposal of a very-large-corpus acquisition method by cell-formed registration. In *Proc. of the 3rd international conference on language resources and evaluation (LREC-2002)*, pages 326–328.

E. Sumita, Y. Akiba, T. Doi, A. Finch, K. Imamura, M. Paul, M. Shimohata, and T. Watanabe. 2003. A corpus-centered approach to spoken language translation. In *Proc. of 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 171–174.

E. Sumita. 2001. Example-based machine translation using DP-matching between word sequences. In *Proc. of the ACL 2001 Workshop on Data-Driven Methods in Machine Translation*, pages 1–8.

T. Takahashi, K. Nawata, K. Inui, and Y. Matsumoto. 2003. Effects of structural matching and paraphrasing in question answering. *IEICE Transactions on Information and Systems*, E86-D(9):1677–1685.

T. Takezawa and G. Kikui. 2003. Collecting machine-translation-aided bilingual dialogues for corpus-based speech translation. In *Eurospeech-2003*, pages 2757–2760.

T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto. 2002. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proc. of the 3rd international conference on language resources and evaluation (LREC-2002)*, pages 147–152.

T. Takezawa. 1999. Building a bilingual travel conversation database for speech translation research. In *Proc. of the 2nd international workshop on East-Asian resources and evaluation conference on language resources and evaluation*, pages 17–20.

Y. Tanaka. 2004. Contrast of the request forms in Japanese and German: How to borrow a pen. *Mathematical Linguistics*, 24(4):193–213. (in Japanese).

The National Institute for Japanese Language, editor. 1964. *Bunrui-Goi-Hyo*. Shuei shuppan. in Japanese.

C. Tillmann, S. Vogel, H. Ney, A. Zubiaga, and H. Sawaf. 1997. Accelerated dp based search for statistical translation. In *Proc. of 5th EUROSPEECH*, pages 2667–2670.

K. Torisawa. 2001. A nearly unsupervised learning method for automatic paraphrasing of Japanese noun phrases. In *Pro. of the Workshop on Automatic Paraphrasing: Theories and Applications*, pages 63–72.

N. Ueffing, K. Macherey, and H. Ney. 2003. Confidence measures for statistical machine translation. In *Proc. of the 9th Machine Translation Summit (MT Summit IX)*, pages 394–401.

W. Wahlster, editor. 2000. *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer.

Alex Waibel. 1996. Interactive translation of conversational speech. *IEEE Computer*, 29(7):41–48.

T. Watanabe and E. Sumita. 2003. Example-based decoding for statistical machine translation. In *Proc. of the 9th MT Summit*, pages 410–417.

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

T. Watanabe, M. Shimohata, and E. Sumita. 2002. Statistical machine translation on paraphrased corpora. In *Proc. of the 3rd international conference on language resources and evaluation (LREC-2002)*, pages 1954–1957.

S. Yamada, E. Sumita, and H. Kashioka. 2000. Translation using information on dialogue participants. In *Proc. of the ANLP-NAACL2000*, pages 37–43.

S. Yamamoto. 2000. Toward speech communications beyond language barrier - research of spoken language translation technologies at ATR -. In *Proc. of International Conference on Spoken Language Processing (ICSLP)*, volume 4, pages 406–411.

K. Yamamoto. 2002. Acquisition of lexical paraphrases from texts. In *Proc. of 2nd International Workshop on Computational Terminology (Computerm 2002)*, pages 22–28.

T. Yoshimi, I. Sata, and Y. Fukumochi. 2000. Automatic preediting of English sentences for a robust English-to-Japanese MT system. *Journal of Natural Language Processing*, 7(4):99–117. (in Japanese).

List of Publications

Journal Papers

1. Mitsuo Shimohata, Eiichiro Sumita, and Yuji Matsumoto. A Method for Retrieving a Similar Sentence and Its Application to Speech Translation (in Japanese), Journal of Natural Language Processing, Vol. 11, No. 4. (to appear)

2. Mitsuo Shimohata, Eiichiro Sumita, and Yuji Matsumoto. Extracting Synonymous Expressions from a Parallel Corpus for Machine Translation (in Japanese), Information Processing Society of Japan (IPSJ) Journal, Vol. 44, No. 11, pp. 2854-2863, 2003.

International Conferences

1. Mitsuo Shimohata, Eiichiro Sumita, and Yuji Matsumoto. Method for Retrieving a Similar Sentence Its Application to Machine Translation, In *Proceedings of 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, 2004. (to appear)

2. Mitsuo Shimohata, Eiichiro Sumita, and Yuji Matsumoto. Building a Paraphrase Corpus for Speech Translation, In *Proceedings of 4th international conference on language resources and evaluation (LREC)*, pp. 453–457, 2004.

3. Mitsuo Shimohata, Eiichiro Sumita, and Yuji Matsumoto. Example-based Rough Translation for Speech-to-Speech Translation, In *Proceedings of the 9th Machine Translation Summit*, pp. 354–361, 2003.

4. Eiichiro Sumita, Yasuhiro Akiba, Takao Doi, Andrew Finch, Kenji Imamura, Michael Paul, Mitsuo Shimohata and Taro Watanabe. A Corpus-Centered Ap-

proach to Spoken Language Translation, In *10th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 171–174, 2003

5. Mitsuo Shimohata, Eiichiro Sumita, and Yuji Matsumoto. Automatic paraphrasing based on parallel corpus for normalization, In *Proceedings of 3rd international conference on language resources and evaluation (LREC)*, pp. 453–457, 2002.

6. Mitsuo Shimohata and Eiichiro Sumita. Converting Morphological Information Using Lexicalized and General Conversion, In *Proceedings of 1st International Conference on Intelligent Text Processing and Computational Linguistics (CICLING)*, pp. 319–331, 2001

Other Publications

1. Mitsuo Shimohata, Eiichiro Sumita, and Yuji Matsumoto. Manual Paraphrasing on Spoken Language for Speech Translation (in Japanese), In *Proceedings of the 10th Meeting of the Association for Natural Language Processing*, pp. 177–180, 2004

2. Mitsuo Shimohata, Toshiyuki Takezawa, and Genichiro Kikui. Construction of English paraphrase Corpus on Travel Conversation and its Analysis (in Japanese), In *Proceedings of the Information Technology Letters*, pp. 83–85, 2003

3. Mitsuo Shimohata, Eiichiro Sumita, and Yuji Matsumoto. Retrieving Meaning-equivalent Sentences for Example-based Rough Translation, In *Proceedings of the HLT-NAACL 2003 Workshop Building and Using Parallel Texts: Data Driven Machine Translation and Beyond the 10th Meeting of the Association for Natural Language Processing*, pp. 50–56, 2003

4. Mitsuo Shimohata and Eiichiro Sumita. Similar Sentence Retrieval for Example-based Rough Translation (in Japanese), In *Proceedings of the 9th Meeting of the Association for Natural Language Processing*, pp. 349–352, 2003

5. Mitsuo Shimohata and Eiichiro Sumita. Acquiring Lexical Paraphrases from a Parallel Corpus (in Japanese), In *Proceedings of the Forum on Information Technology 2002*, pp. 183–184, 2002

6. Mitsuo Shimohata and Eiichiro Sumita. Identifying Synonymous Expressions for Example-based Machine Translation (in Japanese), In *Proceedings of the Information Technology Letters*, pp. 73–74, 2002

7. Kazutaka Takao, Mitsuo Shimohata, Kenji Imamura, and Hideki Kashioka. Comparative Study between Coverage of E-to-J and J-to-E Dictionaries and Its Application (in Japanese), In *Proceedings of the 7th Meeting of the Association for Natural Language Processing*, pp. 58–61, 2001

Award

1. Mitsuo Shimohata. Information Processing Society of Japan (IPSJ) FIT Paper Award, 2002. Identifying Synonymous Expressions for Example-based Machine Translation.