Automatic paraphrasing based on parallel corpus for normalization

Mitsuo Shimohata* **and Eiichiro Sumita***

*ATR Spoken Language Translation Research Laboratories
2-2 Hikaridai Seika-cho Soraku-gun Kyoto 619-0288 Japan
{mitsuo.shimohata eiichiro.sumita}@atr.co.jp

Abstract

There are various ways to express the same meaning in natural language. This diversity causes difficulty in many fields of natural language processing. It can be reduced by normalization of synonymous expressions, which is done by replacing various synonymous expressions with a standard one. In this paper, we propose a method for extracting paraphrases from a parallel corpus automatically and utilizing them for normalization. First, synonymous sentences are grouped by the equivalence of translation. Then, synonymous expressions are extracted by the differences between synonymous sentences. Synonymous expressions contain not only interchangeable words but also surrounding words in order to consider contextual condition. Our method has two advantages: 1) only a parallel corpus is required, and 2) various types of paraphrases can be acquired.

1. Introduction

Natural languages are so expressive that there are various ways to represent the same meaning. This rich expressiveness is extremely useful in human communication. We can convey various additional intentions or nuances by choosing a particular expression.

However, such diversity causes difficulty in natural language processing. If a certain meaning can be expressed in various ways, we have to comprehend this variation to fully grasp this meaning. If we want to retrieve the information of "pamphlet," we have to also search through the information of synonymous words such as "brochure" and "booklet."

"Normalization" is an operation that replaces synonymous expressions with one standard expression. It equalizes the words having basically the same meaning and reduces the variation of synonymous expressions. Although this operation sometimes eliminates nuances and minor information, it is still very useful in many fields.

In this paper, we describe a method of automatic paraphrasing of synonymous expressions for normalization. First, synonymous sentences are grouped by the equivalence of their translations. Then synonymous expressions are acquired from the differences between synonymous sentences. The extraction and filtering of synonymous expressions is built on DP-matching (Cormen et al., 2001). Standard expressions are determined by their frequencies in the corpus. Our method has two advantages: (1) The only required resource is a tagged parallel corpus[1], and no other knowledge or resource, such as sentence structure, a dictionary, or a thesaurus, are used; (2) Various types of paraphrases, such as content words, functional words, and synonyms, can be acquired.

2. Related Works

Paraphrasing is useful for many applications, such as text generation, multi-document summarization, and information retrieval. Automatic acquisition of paraphrases from a corpus is a major approach to the collection of paraphrases. Synonymous multi-word terms can be identified by morphological and syntactic differences (Jacquemin et al., 1997). However, this requires rich linguistic knowledge, and the types of applicable terms are limited.

Various types of paraphrases could be acquired in (Barzilay and McKeown, 2001). Their method is based on the iteration of two processes: extraction of contexts and extraction of interchangeable words. Extracted paraphrases do not include contextual information. They reported that the precision of paraphrases is deeply influenced by context. An evaluation of paraphrases showed that those obtained when taking context into account have a much higher precision than when context is not considered.

Automatic word clustering involves extraction of lexical paraphrases (Langkilde and Knight, 1998), (Lin, 1998), (Frakes and Baeza-Yates, 1992). The types of applicable words, however, are limited to content words, and clustered words are not always interchangeable, such as "dog" and "cat."

The major difference between our method and the above works is our emphasis on applying contextual information. The equivalence of synonymous words is often influenced by the context. The expressions used in our method include the words surrounding synonymous words, which are used as contextual information.

3. Normalization

The operation of "normalization" means that synonymous expressions are replaced with one standard expression. Which expression is most suitable as the standard expression? We use the most frequent expression in the corpus as the standard expression. The operation that replaces the minor expressions with the major expression can delete the minor expressions from the corpus, and it simplifies the corpus. We call this operation normalization in this paper.

The replacement direction from minor to major allows us to paraphrase the expressions having not only an equivalent but also an inclusive semantic relation. For example, the words "picture" and "photo" have an inclusive relation. "Picture" has the meaning of both *photo* and *painting*. The replacement of "photo" with "picture" (A $\rightarrow$ B, C $\rightarrow$ D)

[1] A tagged corpus is one in which morphological analysis has been done.

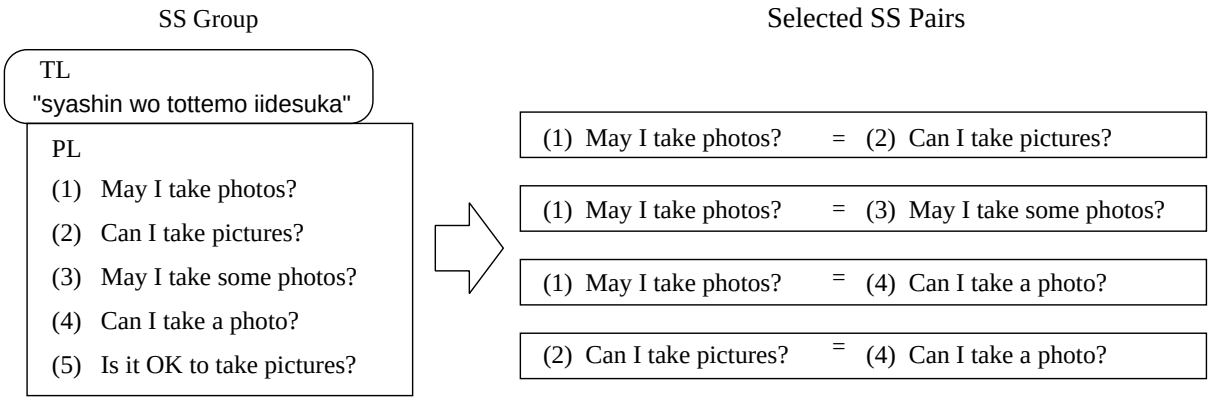

Figure 1: Extraction of SS Pairs

causes little problem, while the inverse replacement sometimes causes a problem, such as (D → C).

(A)	May I take a *photo*?
(B)	May I take a *picture*?
(C)	He drew the *photo* yesterday.
(D)	He drew the *picture* yesterday.

4. Extraction of Synonymous Expressions

The paraphrasing rules described in this paper consist of several synonymous expressions. The most frequent expression is marked as the "major expression," and other expressions are marked as the "minor expressions." Paraphrasing is done by replacing the various minor expressions with the major expression. (Extracted paraphrases are shown in figure 3.) The aim and idea of synonymous expressions are described in section 4.1., and normalization is explained in section 3..

4.1. Basic Idea of Synonymous Expression

The aim of our paraphrasing method is to capture lexical synonymy rather than syntactic. Detection of syntactic synonymy is very difficult since it requires rich information, such as sentence structure and morphological equivalences. We derive lexically synonymous expressions from the differences between sentences that are almost the same. We assume that "almost the same sentences" have lexical differences but no syntactical difference.

The interchangeability of synonymous expressions frequently depends on the context. The words "call" and "phone" are synonymous in the context of telephoning but not synonymous in other contexts. The auxiliary verbs "would" and "could" are interchangeable if they are used in euphemistic request sentences like "(could | would) you pass me the salt?" but not synonymous in other sentences. Therefore, it is necessary to take contextual information into account in determining whether expressions are synonymous. We use the words surrounding synonymous words as contextual information. We use "synonymous expression" (SE) to designate not only synonymous words but also the surrounding words hereafter.

The determination of "almost the same sentences" and the extraction of SE are built on DP-matching. Both minimum edit distance and edit operations (insertion, deletion, substitution) can be computed by DP-matching. Based on the edit operations, the types of SE pairs can be classified into two: substitution and insertion/deletion. Both expressions in substitution pairs have variant words, while one of the expressions in the insertion/deletion pair does not have variant words. We exclude the insertion/deletion pairs for two reasons based on the preliminary experiment: (1) Interjections such as "thank you" and "please" are dominant. These can be easily handled by using dictionaries. (2) Many other words concern context, such as "my," "today's," and "this." It may be safe to preserve them to avoid the confusion of different contexts.

In our method algorithm, sentences are regarded as mere word sequences, and no linguistic information is used for filtering. Therefore, this algorithm can be applied to the tagged text of any language, and the types of paraphrases that can be acquired are not restricted.

4.2. Method

The procedure of generating paraphrasing rules is described below in order of process.

4.2.1. SS Group

"Synonymous sentences" (SS) are defined as sentences that have the same translations in the parallel corpus. The left part of figure 1 shows an example of an SS group. Here, all of the English sentences have the same Japanese translation of "syashin wo tottemo iidesuka." In other words, this Japanese sentence forms an SS group that includes five sentences.

4.2.2. Extraction of SE Pairs

For all pairs of SS, DP-matching is applied. Sentences are regarded as mere word sequences including "head-of-sentence" and "end-of-sentence." Words correspond to elements in DP-matching, and they are identified by their surface form and part-of-speech (POS). Sentence pairs that have more than three variant words in either sentence are excluded. From the SS group in figure 1, ten types of SS pairs can be acquired. Among them, the four sentence pairs

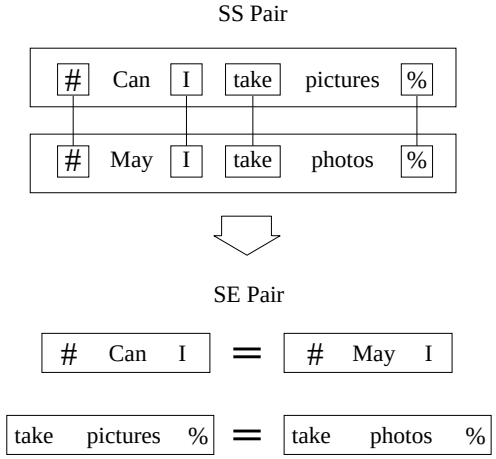

Figure 2: Extraction of SE Pairs

shown in the right side of figure 1 are selected since both sentences in each pair have less than two variant words.

Next, SE pairs consisting of variant words and surrounding common words are extracted. The SE pairs that have no variant words in either expression are excluded. Figure 2 shows an example of SE extraction from DP-matched sentence pairs. Common words are bound by lines, and the symbols "#" and "%" are special marks indicating head-of-sentence and end-of-sentence, respectively. Two SE pairs can be acquired from this sentence pair.

Some SE pairs can be extracted from different sentence pairs in the same group. Each SE pair from a single group is equally counted as one regardless of its frequency. For example, the SE pair "# Can I = # May I" is counted as one, as with other pairs, although it can be extracted from the two combinations (1)-(2) and (1)-(4).

4.2.3. Filtering

Collected SE pairs are filtered by two criteria: frequency and co-occurrence ratio, which is the ratio of the co-occurrence of a minor expression with the major expression to all instances of that minor expression in the corpus. The former involves the case where the synonymy of the pair is valid, and the latter involves the case where the synonymy is invalid. A threshold was set up heuristically, and it is commonly used for English and Japanese. At this stage, the frequency of each expression in the corpus is counted.

Frequency

SE pairs whose frequency is less than three are excluded. The remaining SE pairs must appear in more than three different SS groups.

Co-occurrence Ratio

This filtering criterion removes the SE pairs containing minor expression that have a relatively small co-occurrence with the major expression. These pairs show synonymy that is valid only under a few limited conditions. SE pairs with a co-occurrence ratio of less than 5% are removed. We do not put such a constraint on the major expression. This enables acquisition of pairs having not only equivalent but also inclusion relations.

Table 1: Statistics of the Corpus

	Training		Evaluation	
Sentence (token)	162,319		10,150	
Sentence (type)	97,092 102,406	(E) (J)	8,671 8,922	(E) (J)
Average Length	5.8 6.9	(E) (J)	5.8 6.8	(E) (J)

4.2.4. Synonymous Expression Clusters

SE pairs are clustered by a transitive relation. If A = B and B = C, then A, B, C form the SE cluster. Within the SE cluster, the most frequent expression in the corpus is marked as the major expression.

4.3. Examples of Generated Rules

Examples of extracted paraphrasing rules are shown in figure 3. We refer to the language to be paraphrased as "PL" and to the translation language as "TL." Rules tagged E* are examples of Eng-Jpn [2] rules, and those tagged J* are examples of Jpn-Eng. The major expression in each rule is located at the top.

Rules E1 and J2 describe the exchange of auxiliary verbs. These expressions have basically the same meaning if the nuance is ignored. Rules E5 and E6 describe the equivalence of abbreviations. Rule J1 describes the difference in character types. All expressions in J1 mean "a cold" in English and are pronounced "kaze." This rule says that "kaze" can be written in kanji, katakana, and hiragana[3].

5. Experiment

5.1. Data

We use a bilingual corpus of travel conversation, which has Japanese sentences and corresponding translations in English (Sumita, 2001). Since the translations were made sentence-by-sentence, this corpus was sentence-aligned from its origin. Morphological analysis was carried out with our morphological analysis tools. The corpus was divided into training data and evaluation data by extracting evaluation data randomly from the entire set of data. The statistics of each type of data is shown in table 1.

Sentences consisting of fewer than three words were skipped in our experimental data, since the meaning of short sentences can be widely different by context. For example, "Ticket please" can have the meanings of "I want to buy a ticket" or "Please show your ticket." These short sentences comprise 11% of the English data and 7% of the Japanese data.

[2] We represent combination of PL and TL as "PL-TL." For example, "Eng-Jpn" means that English is the PL and Japanese is the TL.

[3] These are character script types of the Japanese language

E1	#	**Could**	**you**
	#	Would	you
	#	Can	you
	#	Will	you
E2	#	**Nice**	**to**
	#	Glad	to
	#	Pleased	to
	#	Happy	to
E3	**a**	**guarantee**	**%**
	a	warranty	%
E4	**the**	**toilet**	**%**
	the	bathroom	%
	the	lavatory	%
	the	restrooms	%
E5	**what**	**'s**	**wrong**
	what	is	wrong
E6	**I**	**'m**	**a**
	I	am	a
E7	**a**	**bad**	**cough**
	a	terrible	cough
E8	**anything**	**cheaper**	**%**
	anything	less expensive	%
J1	#	Iw<Y	$r
	#	$+$<	$r
	#	%+%<	$r
J2	$$$$	G9	$+
	$$$$	NG$9	$+
	$$$$	G7g&	$+
	$$$$	NG7g$&	$+
J3	#	%H%$%l	$O
	#	$*<j@v$$	$O
	#	2=>Q<<	$O
J4	$N	NA6b	G9
	$N	CMCJ	G9
	$N	6b3[	G9

Figure 3: Examples of Rules

Table 2: Numbers of Rules

PL – TL	Rules	Exp.	E/R
Eng–Jpn	264	583	2.21
Jpn–Eng	423	945	2.23

5.2. Number of Rules

From our trilingual corpus, six combinations of PL and TL can be selected. The numbers of extracted rules, included expressions (Exp.) and included expressions per rule (E/R) by the combination of PL and TL are shown in table 2.

Table 3: Ratios of POS Types

	Eng-Jpn	Jpn-Eng
Noun	26	23
Pronoun	20	2
Verb	25	28
Adjective	5	5
Be-Verb	14	-
Preposition	5	-
Determinative	11	-
Auxiliary Verb	11	21
Particle	-	36

5.3. Types of Rules

Various types of paraphrasing rules can be acquired by our method. The ratios of main POSs to total rules are shown in table 3.

Interestingly, the ratios of content words, such as nouns, verbs, and adjectives, are almost the same independent of PL and TL. On the other hand, functional words, such as pronouns and auxiliary verbs, show obvious differences by PL and TL. For example, many more rules concerning pronouns are generated in using English as PL than in using Japanese as PL.

5.4. Criteria for Evaluation of Paraphrasing

The correctness of paraphrasing was manually evaluated by natives of PL. Paraphrased sentences were evaluated for classification into one of four classes: "Same," "Different" (Dif.), "Semantical Error" (Sem.), and "Syntactical Error" (Syn.). A description of each evaluation class is given below.

Same Paraphrased sentences are natural and represent basically the same meaning as the original.

Dif. Paraphrased sentences are natural but represent different meanings.

Sem. Paraphrased sentences are syntactically proper but semantically improper. No comparison with the original sentence is considered.
ex) "What time does this train land?"

Syn. Paraphrased sentences are syntactically improper.
ex) "Are you have a ticket?"

"Same" is the only evaluation class that satisfies correct paraphrasing, and other evaluations are judged as incorrect paraphrasing. Details of the criteria used in the case of English are given below. The criteria for Japanese follow these guidelines.

1. The available context of the paraphrased sentence sometimes becomes broader or narrower compared with the original sentence. With expansion or reduction of the available context, a sentence is considered "Same." For example, the paraphrased sentence "What time does it start?" has basically the

Table 4: Result of Applied Ratio and Precision

PL – TL	Total	AR	**Same**	Dif.	Sem.	Syn.
Eng–Jpn	7,564	17.1	83.1	4.2	2.1	10.4
Jpn–Eng	8,092	13.9	93.5	1.2	4.9	0.4

same meaning as the original sentence "When does it start?", although the available context of the paraphrased sentence is narrower.

2. The degree of politeness is excluded from the basic meaning. The differences among "Would you," "Could you," and "Won't you" do not affect the basic meaning.

3. Differences in the pronouns used do not affect the basic meaning. The interchange of "this," "that," and "it" does not affect the basic meaning. However, the exchange of personal and non-personal pronouns is judged as semantically improper. The paraphrased sentence "It wants to buy a ticket" is semantically improper.

5.5. Applied Ratio and Precision of Paraphrasing

We evaluated the precision of two types of paraphrasing rules, Eng-Jpn and Jpn-Eng. The results of applied ratio (AR) and precision are shown in table 4.

There were many syntactical errors in English paraphrasing but relatively few in Japanese paraphrasing. This difference was due to the respective strengths of the syntactical constraints in English and Japanese. Since the area of each paraphrasing rule is partial, syntactical discrepancies between paraphrased and other parts can be caused. Examples of rules that often cause syntactical errors are shown below.

"Where are you" = "Where do you"
"Where are the " = "Where is the"

6. Conclusions

In this paper, we described a method for acquiring paraphrasing rules from a parallel corpus. These rules normalize the expressions of the corpus by replacing synonymous expressions with a major expression. Extraction and filtering of expressions are based on DP-matching, so no other resources are required, and a wide range of paraphrases can be acquired. The effectiveness of our method was demonstrated by an experiment using English and Japanese as paraphrased languages.

We have two plans for the future. One is to improve the method by adjusting the length of the surrounding context. Longer context would constrain the application and improve precision. Shorter context would extend the applied ratio. Proper adjustment of the context length could improve both applied ratio and precision. The other plan is to apply our method to other NLP fields, such as example-based machine translation, statistical machine translation, and information retrieval. Our paraphrasing method could have a great effect in these fields.

7. Acknowledgment

The research reported here was supported in part by a contract with the Telecommunications Advancement Organization of Japan entitled, "A study of speech dialogue translation technology based on a large corpus".

8. References

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

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

W. B. Frakes and R. Baeza-Yates, editors, 1992. *Information Retrieval Data Structures & Algorithms*, chapter Thesaurus Construction, pages 161–218. Prentice Hall.

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*.

I. Langkilde and K. Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *COLING-ACL*, pages 704–710.

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

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