Identifying Synonymous Expressions from a Bilingual Corpus for Example-Based Machine Translation

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

Example-based machine translation (EBMT) is based on a bilingual corpus. In EBMT, sentences similar to an input sentence are retrieved from a bilingual corpus and then output is generated from translations of similar sentences. Therefore, a similarity measure between the input sentence and each sentence in the bilingual corpus is important for EBMT. If some similar sentences are missed from retrieval, the quality of translations drops. In this paper, we describe a method to acquire synonymous expressions from a bilingual corpus and utilize them to expand retrieval of similar sentences. Synonymous expressions are acquired from differences in synonymous sentences. Synonymous sentences are clustered by the equivalence of translations. Our method has the advantage of not relying on rich linguistic knowledge, such as sentence structure and dictionaries. We demonstrate the effect on applying our method to a simple EBMT.

1 Introduction

Example-based machine translation (EBMT) is one of the main approaches to corpus-based machine translation, and it offers the advantage of requiring far less manual work than rule-based machine translation. The basic idea of EBMT is that the translation of an input sentence can be acquired by modifying translations of similar sentences, as is done in the human translation process (Nagao, 1981). Therefore, the selection of similar sentences from a bilingual corpus is important for EBMT. A similarity measure with low retrieval ability lessens the exploitation of the bilingual corpus and results in bad or no translation.

Identifying synonymous expressions is an effective way to expand the retrieval of similar sentences. It equates different expressions that have almost the same meaning and shortens the distance between sentences that are essentially the same but look different.

In this paper, we describe a method to extract synonymous expressions from a bilingual corpus. Extraction is based on differences between synonymous sentences by dynamic programming match (DP-match)(Cormen et al., 2001). The method has the advantage that it does not require rich linguistic knowledge, such as sentence structure and dictionaries.

2 Basic Idea

The synonymous expressions (SE) defined in this paper are focused on lexical variations. SE are extracted by comparing synonymous sentences (SS). In this section, we describe the basic idea of SS and SE.

2.1 Synonymous Sentences (SS)

SS are defined as sentences that they have same basic meaning and lexical differences. Satisfaction of both conditions can be verified by a bilingual corpus.

The condition having the same basic meaning can be verified by the equivalence of translations. The left side of figure 1 shows an example of a sentence group that has the common Japanese translation "syashin wo tottemo iidesuka." The sentences in this sentence group satisfy the condition having same basic meaning.

The condition having lexical variations is verified by an edit distance of DP-match between two sentences, which represents the number of word-level differences. Sentence pairs with small edit distances share many common words and are considered to have the same structure. The right side of figure 1 shows the SS group,

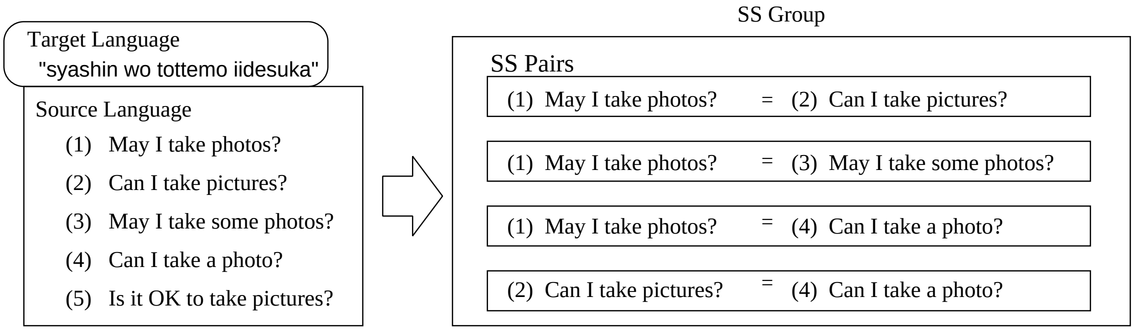
Target Language
"syashin wo tottemo iidesuka"
Source Language
(1) May I take photos?
(2) Can I take pictures?
(3) May I take some photos?
(4) Can I take a photo?
(5) Is it OK to take pictures?
SS Group
SS Pairs
(1) May I take photos? = (2) Can I take pictures?
(1) May I take photos? = (3) May I take some photos?
(1) May I take photos? = (4) Can I take a photo?
(2) Can I take pictures? = (4) Can I take a photo?

Figure 1: Extraction of SS Pairs

which consists of SS pairs, derived from the sentence group of the left side of figure 1. These SS pairs have the same structure and lexical differences. Sentence 5 has a large edit distance compared with other sentences since it has a different structure. Consequently, it is excluded from the SS group shown in the right side of figure 1.

2.2 Synonymous Expressions (SE)

SE have three features: (1) they include surrounding words of different expressions as contextual conditions, (2) they have influence on the target language, and (3) they are not restricted to any type of variation, such as content words or functional words. Details of each feature are described in the following.

2.2.1 Contextual Condition

Since in many cases the synonymy of expressions depends on the context, SE must have contextual conditions. The words “picture” and “photo” are synonymous if “picture” denotes the meaning of “*photo*,” while they are not synonymous if “picture” denotes the meaning of “*painting*.” 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 are not synonymous in other sentences.

The common words surrounding different words are used as a contextual condition. They have the advantage that they are effective enough as a contextual condition and are easy to acquire. For example, the expressions “take pictures” and “take photos” are synonymous in most cases. The same applies to the expressions “#[1] Would you” and “# Could you.”

An example of SE extracted from a corpus of travel conversation (Takezawa et al., 2002) (detail is described in 4.1) are shown in figure 2. The clusters tagged with E* represent English SE clusters based on Japanese translation, and those tagged with J* represent Japanese SE clusters based on English. The surrounding words of E1 properly work as a contextual condition. Unfortunately, many SE have unnecessary conditions or need other contextual conditions.

2.2.2 Influence on Target Language

Extracted SE have influence on the target language, since the synonymy of SE depends on the equivalence of translations in the target language. It is important that the influences are quite valuable to similarity measure in translingual application. Though some of them seem inappropriate from the viewpoint of source language alone, they have no bad influence on the similarity measure.

In other words, differences in such SE show that they are not distinguished from the viewpoint of the target language. Figure 3 shows English SE with Japanese influences. C1 equates the gender of a person, and C2 equates the difference of singular/plural. These differences are seldom expressed in Japanese. C3 equates similar but different objects. They share the same translation “saifu” in Japanese.

2.2.3 Unrestricted to Types

SE do not have restrictions on the types of differences. Therefore, there are many types of differences in SE. The extracted SE shown in

[1] This represents “start-of-sentence”

E1	#	Could	you
	#	Would	you
	#	Can	you
	#	Will	you
E2	a	guarantee	%[2]
	a	warranty	%
E3	the	toilet	%
	the	bathroom	%
	the	lavatory	%
	the	restrooms	%
E4	what	's	wrong
	what	is	wrong
E5	a	bad	cough
	a	terrible	cough
J1	itai	desu	%
	itai	ndesu	%
	itai	nodesu	%
J2	i i	desu	ka
	i i	nodesu	ka
	i i	desyou	ka
	i i	nodesyou	ka
J3	#	toire	wa
	#	otearai	wa
	#	kesyoushitsu	wa
J4	no	ryoukin	desu
	no	nedan	desu
	no	kingaku	desu

Figure 2: Examples of SE Clusters

C1	tell	him	to
	tell	her	to
C2	emergency	exit	%
	emergency	exits	%
C3	my	wallet	%
	my	purse	%

Figure 3: English SE with Japanese influences

4.1 vary in many types. The SE of E1 differ in degree of politeness, those of E2 and E3 differ in synonyms, and the SE of E4 differ in abbreviation.

3 Procedure for Extraction of SE

The procedure to extract clusters of SE, in which all expressions are synonymous, is described below. The following description is based on extraction of English SE from Japanese translations. Japanese SE from English translations can be extracted the same way. A bilingual corpus is expressed as a set of sentence pairs {(E1=J1), (E2=J2), ... , (En=Jn)}. Some of the sentences are equal (e.g. E4 = E9, J1 = J5 = J11).

3.1 Clustering SS Group

English sentence groups that share the same Japanese translation are clustered. If J1 = J5 = J9, then English sentence group {E1,E5,E9} is clustered.

Each group is tested as described below:

1. All combinations of sentence pairs are extracted.
2. Apply DP-match to sentence pairs, regarding sentences as word-sequences including "head-of-sentence" and "end-of-sentence." Words are identified by their surface form and part-of-speech (POS). Results of edit operations extracted by the DP-match are preserved for the following step.
3. Select both sentence pairs if their edit distance is within two.

Selected groups are recognized as SS groups.

3.2 Extraction of SE Pairs

SE are based on edit operations extracted from SS pairs. SE include not only differences but also common words surrounding those differences. Figure 4 shows the extraction of SE pairs from SS pairs. DP-match is applied to the SS pair, and corresponding words are bound. Two SE pairs "# Can I"="# May I" and "take pictures %"= "take photos %" can be extracted.

The frequency of each SE pair is also counted for the following process. These frequencies are based on SS groups, namely, counting the number of SS groups in which the target SE appears. For example, the SE pair "# Can I" = "# May I" can be extracted from the SS group shown in figure 1. Though this SE pair can be extracted from two SS pairs (1)-(2) and (1)-(4), it adds only one frequency to this SE pair counter.

[2] This represents "end-of-sentence"

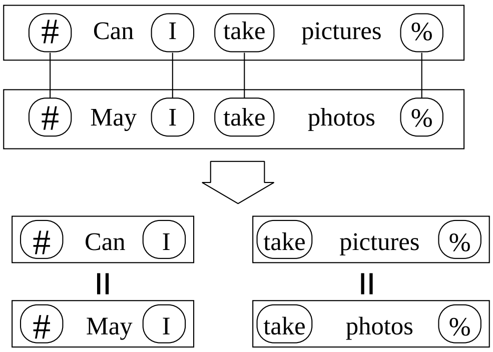

Figure 4: Extraction of SE pairs

3.3 Filtering

The collected SE pairs are filtered by two criteria: overlap of SS group and frequency of SE.

Overlap of SS Group
SE pairs in which component expressions have a small overlap are excluded, since it shows that component expressions are synonymous only in restricted cases. The filtering is done by comparing the frequency of SE pairs and that of the more infrequent component expression. If the frequency of the SE pair is lower than five percent of the frequency of the more infrequent component expression, the SE pair is excluded.

Frequency of SE
SE with small frequency are excluded. In this paper, SE occurring only once are excluded. This threshold is commonly used in SE extraction of English and Japanese.

3.4 Clustering Synonymous Expression Pairs into Clusters

SE pairs are clustered by a transitive relation. If Exp1 = Exp2, and Exp2 = Exp3, then Exp1, Exp2, and Exp3 compose the SE cluster.

4 Experiment

We have conducted an experiment using a bilingual corpus of Japanese and English. The effect of our method is demonstrated by comparing the results of two EBMT systems: EBMT without our method (“w/o”) and EBMT with our method (“with”). The architecture of EBMT used for the experiment makes it simple to see the effect of our method, which was evaluated by two criteria: expansion of coverage and quality of translation.

Table 1: Statistics of the Corpus

	Training		Evaluation	
Sentence (Token)	162,319		10,150	
Sentence (Type)	97092 102,406	(E) (J)	8,671 8,922	(E) (J)
Average Length (Words)	5.8 6.9	(E) (J)	5.8 6.8	(E) (J)

Table 2: Number of Extracted Clusters

Source Language	English	Japanese
Clst.	673	912
Expr.	1,512	2,130
Expr. / Clst.	2.2	2.3

4.1 Data

We used a bilingual corpus of travel conversation, which has Japanese sentences and their English translations. This corpus was sentence-aligned, and a morphological analysis was done on both languages by our morphological analysis tools.

The bilingual corpus was divided into training data and evaluation data by extracting evaluation data randomly from the whole set of data. The training data were used to extract SE clusters and the bilingual corpus of EBMT. The evaluation data were used as a set of input sentences. The statistics of the both data sets are shown in table 1.

Sentences consisting of fewer than three words were excluded from the experimental data, since short sentences can represent various meanings according to context.

4.2 Implementation

SE clusters were extracted in both languages from the training corpus. The parameter for extraction was the same in both languages. The numbers of extracted SE clusters (Clst.) and contained expressions (Expr.) in each language are shown in table 2.

The variation in types of extracted SE clusters is shown in 3. Counts of major part-

Table 4: Coverage

	Acceptable Source Sentences			Retrieved Translation		
	w/o	with	**Exp.**	w/o	with	**Exp.**
E to J	2,845	3,034	6.6%	9,666	16,178	67.3%
J to E	3,198	3,419	6.9%	8,966	15,915	77.5%

Table 3: Types of SE by Major POS

POS	English	Japanese
Noun	177	378
Verb	137	219
Pronoun	105	57
Auxiliary verb	77	122
Adverb	38	29
Adjective	42	38

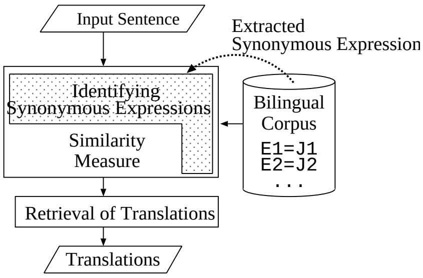

Figure 5: Architecture of EBMT

of-speech (POS) are shown for English and Japanese SE. As described in section 2.2.3, acquired SE are various in types of POS.

The architecture of EBMT used in the experiment is shown in figure 5. The similarity measure module contains our method (dotted box in figure 1). Similarity of the two sentences is measured by an exact-match. It returns only two values: exact-match or not. Translations of exact-match sentences in the corpus are retrieved and output.

The system retrieves sentences that are the same as the input sentence from a bilingual corpus. Then, it outputs translations of retrieved sentences. In the case of the "with" EBMT system, SE between the input sentence and sentences in the corpus were equated.

Translation was done in two directions: from English to Japanese (E to J) and from Japanese to English (J to E). English SE clusters were used for E to J, and Japanese SE clusters were used for J to E.

4.3 Expansion of Coverage

The effect of coverage expansion is divided into two types of input sentences: those acceptable only to "with" EBMT and those acceptable to both "w/o" and "with" EBMT. The former is evaluated by the expansion of acceptable input sentences. The latter is evaluated by the expansion of retrieved translations. Expansion of retrieved translations is useful since many EBMT systems (Sumita, 2001) (Veale and Way, 1997) (Carl, 1999) (Brown, 2000) utilize plural translations from similar sentences to acquire output translation.

The results of the two types are shown in table 4. "Exp." denotes the expansion ratio of with to w/o. The results show an obvious effect on the expansion of the coverage. Interestingly, the expansion effect is similar in English and Japanese.

4.4 Quality of Translations

The quality of translation was evaluated by native speakers of the target languages. They evaluated translations as correct (Cor.) or not. "Cor." means that the translation is basically appropriate for the translation of input sentence. Small differences, such as degree of politeness and exchange of pronouns, are not considered.

Translations of a part of the evaluation data, 1,048 source sentences in E to J and 1,094 in J to E, were evaluated. When an input sentence had plural translation candidates, they were individually evaluated and counted as correct or

Table 5: Accuracy

		w/o	with
J to E	Total	1,552	961
	Cor.	1,395	871
	Acc.	89.9%	90.6%
E to J	Total	1,645	1,055
	Cor.	1,606	1,030
	Acc.	97.6%	97.6%

not. The results are shown in table 5. Sentences of “w/o” represent retrieved sentences by “w/o” EBMT, and that of “with” represent additional retrieved sentences by “with” EBMT. Accuracy (Acc.) denotes the ratio of correct sentences to the total. These results demonstrate that translation qualities of “w/o” and “with” are equivalent for each translation direction.

5 Related Work

Many works have attempted to improve the similarity measure on the lexical level. They require other linguistic knowledge, while our method does not.

A thesaurus has been utilized to measure the semantic distance of words (Sumita, 2001). Semantic distance is proportional to a hierarchical difference between two words. Morphological knowledge and POS have also proven useful (Nirenburg et al., 1994). Weights for the similarity measure are changed by type of word, i.e., content word or functional word.

6 Conclusions and Future Work

In this paper, we described a method to acquire synonymous expressions from a bilingual corpus. The method has the advantage of not requiring rich linguistic knowledge for extraction. The synonymous expressions defined in this paper have three features: (1) they use the words surrounding different words as contextual conditions, (2) they contain the influence of the target language, and (3) they include various types of expressions.

The experiment demonstrates that our method expands the coverage of EBMT without deterioration of translation quality. Furthermore, our method has an equivalent effect on both translation directions, E to J and J to E.

Recently, we have been conducting experiments to investigate the effects under various source/target languages. Volume and validity of extracted synonymous expressions depend on source/target languages. Detailed analysis of the relation would be an interesting future work.

Acknowledgements

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”.

References

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 Machine Translation Summit VII*, pages 250–258.

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

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

S. Nirenburg, S. Beale, and C. Domashnev. 1994. A full-text experiment in example-based machine translation. In *Proc. of the International Conference on New Methods in Language Processing*, pages 78–87.

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.

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 *LREC-2002*.

T. Veale and A. Way. 1997. Gaijin: A bootstrapping, template-driven approach to example-based MT. In *Proc. of the NeMNLP97*.