Method for Retrieving a Similar Sentence and Its Application to Machine Translation

Mitsuo Shimohata†‡ Eiichiro Sumita† Yuji Matsumoto‡
†ATR Spoken Language Translation Research Laboratories
2-2-2 Hikari-dai, Seika-cho, Soraku-gun, Kyoto, 619-0237, JAPAN
and
‡Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0101, JAPAN
{mitsuo.shimohata, eiichiro.sumita}@atr.jp matsu@is.aist-nara.ac.jp

Abstract

In this paper, we propose incorporating similar sentence retrieval in machine translation to improve the translation of *hard-to-translate* input sentences. If a given input sentence is hard to translate, a sentence similar to the input sentence is retrieved from a monolingual corpus of translatable sentences and then provided to the MT system instead of the original sentence. This method is advantageous in that it relies only on a monolingual corpus. The similarity between an input sentence and each sentence in the corpus is determined from the ratio of the common N-gram. We use two conditions to improve the retrieval precision and add a filtering method to avoid inappropriate sentences. An experiment using a Japanese-to-English MT system in a travel conversation domain proves that our method improves the translation quality of hard-to-translate input sentences by 9.8 %.

1 Introduction

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 & Gdaniec 2001) provided a guideline for the manual pre-editing of input sentences. (Mitamura & 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 & 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.

In this paper, 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 1 shows an overview of our proposal. An input sentence

[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.

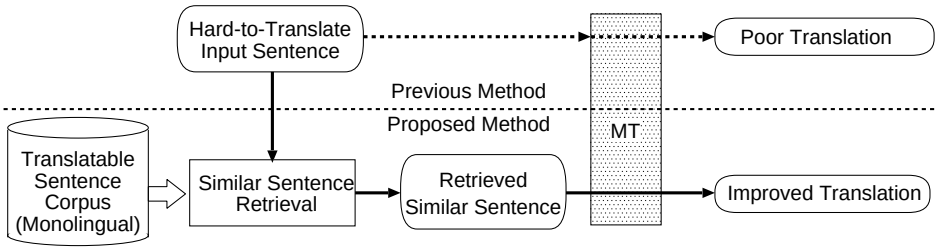

Figure 1: Improving Translation by Similar Sentence Retrieval

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 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 3) Furthermore, we add two additional conditions to improve retrieval precision. (Section 4) We describe an experiment on applying similar sentence retrieval to a Japanese-to-English MT system in Section 5.

2 Experimental Data

We focused on travel conversation, which is a major target domain in speech translation (Sumita et al. 1999) (Wahlster 2000). We used two types of Japanese corpora: a corpus based on machine-aided dialog (MAD) and the basic travel expression corpus (BTEC). The MAD corpus was used as a collection of inputs and the BTEC as a candidate corpus.

The MAD corpus contains 2,135 different sentences. These sentences are a collection of transcribed utterances that were spoken in several travel situations. Therefore, the corpus reflects the characteristics of spoken language. The sentences are divided into two parts: one containing 437 sentences and the other 1,698 sentences. The former part was used in

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

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

an experiment to find a method for similar sentence retrieval (Sections 3 and 4). The latter part was used in an experiment applying similar sentence retrieval to MT (Section 5). The BTEC is a collection of edited colloquial travel expressions often found in phrasebooks. It contains 116,773 different sentences.

The MAD corpus is derived from transcribed utterances, while the BTEC is not. Although both corpora contain expressions frequently used in travel conversation, they have different characteristics in terms of average number of words in a sentence [4] and perplexity (Takezawa et al. 2002). We used the BTEC as the candidate corpus because its sentences are shorter and more grammatical than those in the MAD corpus. Consequently, these sentences are more suitable for application to MT.

3 Method for Measuring Similarity between Two Sentences

Our method for measuring similarity is based on research into automatic evaluation of MT results. Similarity measurement in automatic evaluation only depends on reference sentences[5] and does not require any other knowledge. In addition, many research attempts have demonstrated that automatic evaluation is strongly correlated to 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.

3.1 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 a higher score as it shares more common parts with several reference translations. There are three basic methods for measuring similarity between a test translation and reference translations: the common N-gram, the word error rate (WER), and the 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). WER is a length-normalized Levenshtein distance. This metric is widely used to measure speech recognition errors. 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] MAD: 10.2 words/sentence BTEC: 6.9 words/sentence

[5] These are a set of different, proper translations of the test source sentence.

3.2 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.

$$\text{Precision} = \frac{\text{Common elements}}{\text{Elements in candidate}} \qquad \text{Recall} = \frac{\text{Common elements}}{\text{Elements in input}} \qquad \text{F-measure} = \frac{2PR}{P+R}$$

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 considers word order through the sentence. The common word set method does not take word order information into account.

Common N-gram

Precision in the common N-gram method is equal to that of BLEU (Papineni et al. 2002), which is a precision-based metric. Recall in the common N-gram method is determined by exchanging the role of the candidate sentence in the precision equation for that of an input sentence. The BLEU method uses a brevity penalty, which penalizes short-length sentences, instead of recall because it is difficult to determine the recall value from multiple reference sentences.

The other difference between the common N-gram method and the BLEU method is that our method uses only unigrams and bigrams, while the BLEU method uses unigrams to 4-grams. The requirement in which a retrieved sentence needs to share any 4-gram is too strict for our target domain, since sentences in travel conversation are relatively short.

Common Word Sequence

This method is based on the longest common word sequences by using DP-matching. 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 performance of the common word sequence metric is nearly equivalent to that of the WER metric.

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.

3.3 Additional Conditions

3.3.1 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 content 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.

3.3.2 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 of function word expressions. Expressions consisting of auxiliary verbs and particles show 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 for 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 to 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 the 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.

3.4 Comparing Precision of Each Method and Additional Conditions

3.4.1 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.

Table 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%

3.4.2 Definition of Similar Sentence Retrieval Precision

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

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

A "similar retrieved sentence" is defined as a substitutive sentence that allows a conversation to proceed.

3.4.3 Results

Precisions for combinations of each basic method and additional conditions are shown in Table 1. The conditions described in Sections 3.3.1 and 3.3.2 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 1 indicates a large effect by excluding additional conditions. This improved precision by at least 8% for every basic method. 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.

4 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: the number of missing content words and the number of common content words.

4.1 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.

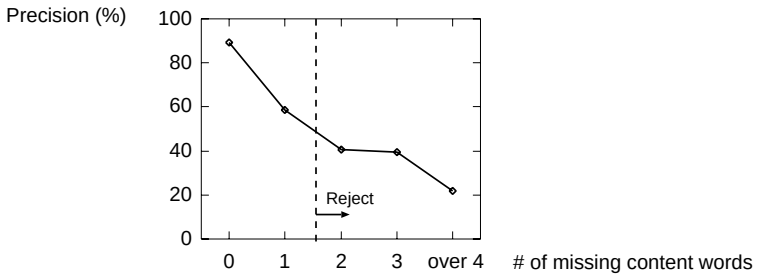

Figure 2: Precision by Number of Missing Content Words

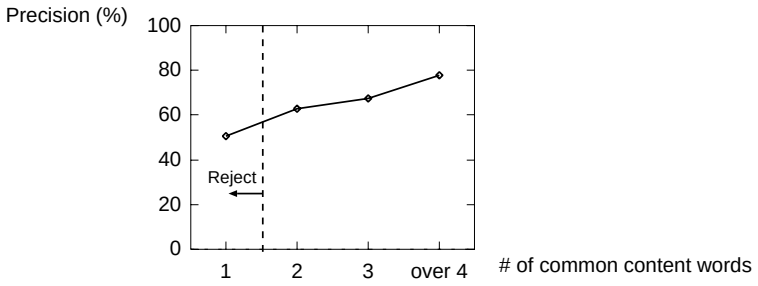

Figure 3: Accuracy by Number of Common Content Words

Figure 2 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 the number of missing content words increased, precision decreased. We determined that a retrieved sentence should remain if it misses fewer than two content words. The precision with one missing content word was 58.8%.

4.2 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 3 shows precision by the number of common content words. As the number of common content words increased, precision also increased. We determined that a retrieved sentence should remain 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 of the condition determined in the previous section.

4.3 Results with Filtering

Figure 4 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 sen-

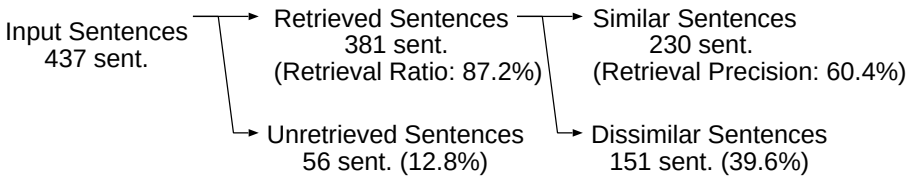

Figure 4: Retrieval Precision with Filtering

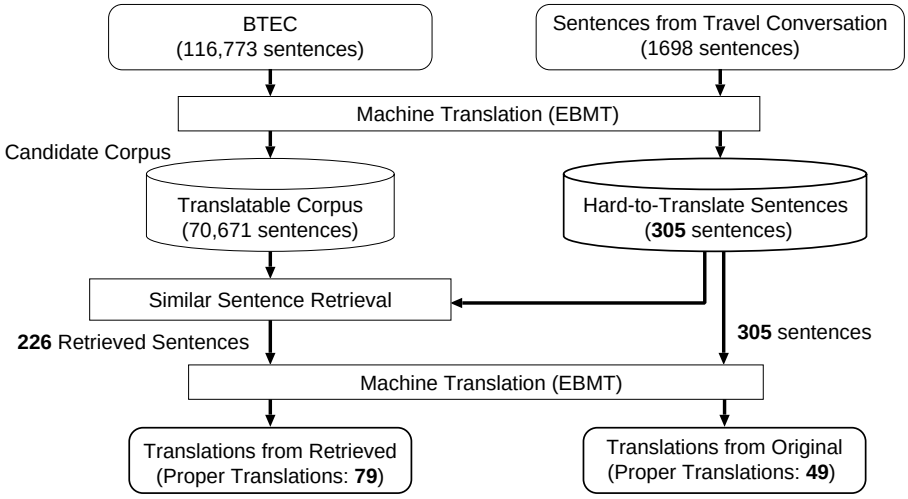

Figure 5: Overview of Experiment

tences 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 determined in Sections 4.1 and 4.2.

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 the coincidence of content words between input and candidate sentences provides at least 77.3% precision regardless of function words.

5 Experiment on Application to Machine Translation

Figure 5 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.

Table 2: Translation Quality of Original and Retrieved Sentences

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

We compared the similarity between input sentences and retrieved sentences manually. The evaluation criterion was the same as that described in Section 3.4.2. 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 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 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%.

5.1 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 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.

5.2 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 4.

6 Conclusion

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%.

Acknowledgment

The research reported here was supported in part by a contract with the National Institute of Information and Communications Technology entitled "A study of speech dialogue translation technology based on a large corpus".

References

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

Doi, T. & E. Sumita: 2003, 'Input sentence splitting and translating', *Proc. of Workshop on Building and Using Parallel Texts, HLT-NAACL 2003*, pp. 104–110.

Mitamura, T. & E. Nyberg: 2001, 'Automatic rewriteng for controlled language translation', *Proc. of NLPRS-2001 Workshop on Automatic Paraphrasing: Theories and Applications*, pp. 1–12.

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

Papineni, K., S. Roukos, T. Ward & 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)*, pp. 311–318.

Sumita, E.: 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*, pp. 1–8.

Sumita, E., S. Yamada, K. Yamamoto, M. Paul, H. Kashioka, K. Ishikawa & S. Shirai: 1999, 'Solutions to problems inherent in spoken language translation: The ATR-MATRIX approach', in *Proc. of the 7th MT Summit*, pp. 229–235.

Takezawa, T., E. Sumita, F. Sugaya, H. Yamamoto & S. Yamamoto: 2002, 'Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world', in *Proc. of the 3rd LREC-2002*, pp. 147–152.

Tillmann, C., S. Vogel, H. Ney, A. Zubiaga & H. Sawaf: 1997, 'Accelerated DP based search for statistical translation', in *Proc. of 5th EUROSPEECH*, pp. 2667–2670.

Ueffing, N., K. Macherey & H. Ney: 2003, 'Confidence measures for statistical machine translation', in *Proc. of the 9th MT Summit*, pp. 394–401.

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