Building a Paraphrase Corpus for Speech Translation

Mitsuo Shimohata[†‡], Eiichiro Sumita[†], and Yuji Matsumoto[‡]

[†]ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai Seika-cho Soraku-gun Kyoto 619-0288 Japan

[‡]Graduate School of Information Science Nara Institute of Science and Technology
8916-5 Takayama Ikoma Nara 630-0101

{mitsuo.shimohata eiichiro.sumita}@atr.jp, matsu@is.aist-nara.ac.jp

Abstract

When a machine translation (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 often seen in spoken language. To reduce these difficulties, we are developing methods to paraphrase input sentences into more translatable ones. In this paper, we report a preliminary Japanese paraphrase corpus. 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 is applied to long sentences, where long sentences are converted into one or several short sentences. We also report a comparison of machine translation quality between the original sentences and the paraphrased sentences. We use two corpus-based machine translation systems in the experiment.

1. Introduction

Machine translation (MT) for spoken language has been developed as a component of speech-to-speech translation. C-star[1], Verbmobil(Wahlster, 2000), and Nespole!(Metze et al., 2002) are well-known projects in this field.

When a machine translation system receives input sentences of spoken language, the following two types of input sentences are difficult to translate: (1) Long input sentences and (2) Input sentences having redundant expressions often seen in spoken language.

We are developing a method for paraphrasing input sentences to reduce this difficulty. A paraphrased sentence shares its main meaning with the original input sentence but is easier to translate. MT performance can be improved by using a paraphrased sentence instead of the original input sentence. In this paper, we report a Japanese paraphrase corpus in which paraphrasing has done by humans. The corpus consists of original sentences derived from travel conversation and paraphrased versions of these sentences. Three paraphrasing methods are used in the corpus: plain, segment, and summary. Plain paraphrasing is applied to short input sentences, and segment and summary paraphrasing are applied to long input sentences.

We report a comparison of machine translation performance between original sentences and paraphrased sentences. Two corpus-based machine translation systems are used in the experiment. We also report the statistical characteristics of these sentences in terms of perplexity.

2. Basic Idea of Paraphrasing

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

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 proper one from 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 a metric of sentence length and the threshold to determine short or long in the following sections.

2.1. Sentence Length Metric

We use "number of content words" as a metric of sentence length. This means that a unit of the metric is a word, and function words are excluded from the counting. 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 of part-of-speech. Content words are defined to include nouns, verbs, adjectives, adverbs, and numerals. Function words are defined to include particles, auxiliary

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

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

Figure 2: Example of Paraphrasing

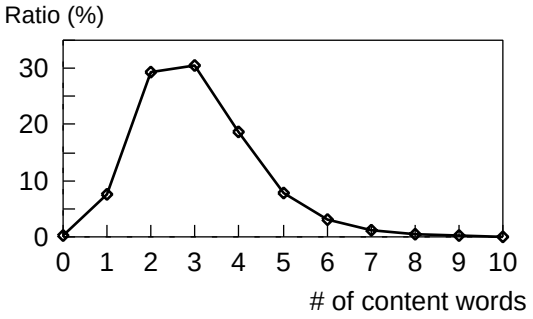

Figure 1: Ratio of Content Words in BTEC Corpus

verbs, and the copula. Compound words, in the case of English "New York," "get off," and "two hundred dollars," are treated as one word.

2.2. 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 more than five content words. Figure 1 shows the ratio of content words in the BTEC corpus.

3. Paraphrasing Methods

In this section, we describe a detail of three paraphrasing methods. Examples of input sentences and paraphrased sentences are shown in Figure 2. To facilitate reader's comprehension, the examples are shown in English and content words are underlined.

3.1. Segment Paraphrasing

A long sentence is divided into several short sentences in the segment paraphrasing method. If some words need to be complemented to make sentences, adding these words is allowed. If it is difficult to divide a given sentence into parts that are all "sentences," the use of phrases is allowed. In the first example in Figure 2, the original sentence includes five content words. It is paraphrased into two short sentences, each of which includes fewer than five content words.

3.2. 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 2, the number of content words is reduced from five to three. Deleted information such as "just" and "computer" are insignificant.

3.3. 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 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.

4. Experiment

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

Figure 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. 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 4.2.). The characteristic of original and paraphrased sentences is analyzed using perplexity (Section 4.3.).

4.1. Experimental MT Systems

Two corpus-based MT systems are 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 sentences from a parallel corpus and modifies the translation of similar sentences to generate

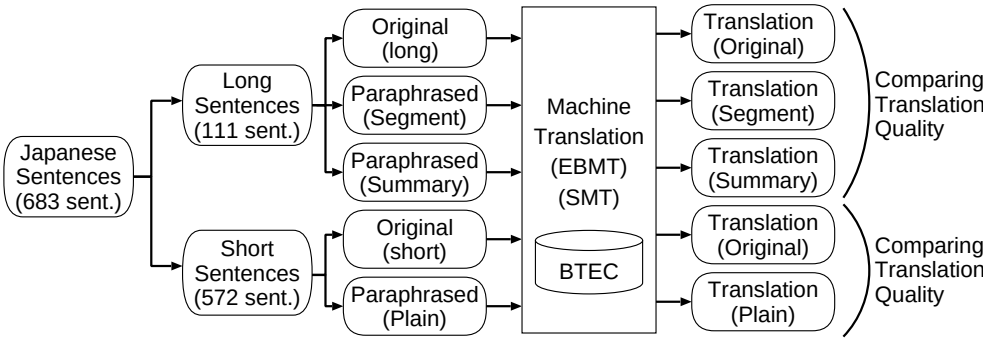

Figure 3: Overview of the Experiment

Table 1: Translation Quality

Sentence Length	Translation Method	Paraphrasing Method	Ratio of Evaluation (%) A	B	C	D	N	Translation Accuracy (%)
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%

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.

4.2. 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 that case, we give an "N" rank.

Table 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 particular, the paraphrasing methods improve the 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.

Table 2: Cross-perplexity

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

4.3. Cross Perplexity

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

Table 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 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%

4.4. Positive/Negative Paraphrasing between Original and Paraphrased Sentences

In this section, we discuss the result of MT quality movement in detail. Table 3 shows a number of positive/negative cases between the original and paraphrased sentences. The MT quality comparison of paraphrased (Para.) and original (Org.) sentences is based on the ranks of A, B, C, D, and N.

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 small negative cases and showed large improvement. However, the SMT system had large negative cases and showed little improvement.

This result suggests that the paraphrasing effect can be improved by eliminating negative paraphrasing. We consider that the following two works are useful for this elimination. Shimohata(Shimohata and Sumita, 2002) proposed a method for extracting local paraphrases from two sentences sharing the same meaning. We can obtain local paraphrases by giving original and paraphrased sentences to this method. Imamura(Imamura et al., 2003) proposed a filtering method of translation rules by automatic evaluation of machine translation. Local paraphrases that are effective to target MT systems can be filtered by this method. The combinational use of the approaches remain our future work.

5. Conclusions and Future Work

We are developing a method for paraphrasing input sentences to facilitate machine translation. In this paper, we reported a Japanese paraphrase corpus as a pilot study. The corpus consists of original sentences derived from a travel conversation corpus and their paraphrased versions. We use 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 and convert long sentences into one of several short sentences.

Experimental result suggests that this paraphrasing strategy has a large effect on EBMT in long sentences but a small effect on SMT in long sentences; Paraphrasing has a small effect on both MTs in 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.

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”.

6. References

Imamura, Kenji, Eiichiro Sumita, and Yuji Matsumoto, 2003. Feedback cleaning of machine translation rules using automatic evaluation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*.

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

Metze, F., 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)*.

Shimohata, Mitsuo and Eiichiro Sumita, 2002. Identifying synonymous expressions from a bilingual corpus for example-based machine translation. In *Proc. of the 19th COLING Workshop on Machine Translation in Asia*.

Sumita, E., 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.*

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

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

Watanabe, T. and E. Sumita, 2003. Example-based decoding for statistical machine translation. In *Proc. of the 9th Machine Translation Summit*.