HLT-NAACL 2003 Workshop: Building and Using Parallel Texts
Data Driven Machine Translation and Beyond , pp. 50-56
Edmonton, May-June 2003

Retrieving Meaning-equivalent Sentences for Example-based Rough Translation

Mitsuo Shimohata **Eiichiro Sumita**
ATR Spoken Language Translation
Research Laboratories
mitsuo.shimohata@atr.co.jp
eiichiro.sumita@atr.co.jp

Yuji Matsumoto
Nara Institute of
Science and Technology
matsu@is.aist-nara.ac.jp

Abstract

Example-based machine translation (EBMT) is a promising translation method for speech-to-speech translation because of its robustness. It retrieves example sentences similar to the input and adjusts their translations to obtain the output. However, it has problems in that the performance degrades when input sentences are long and when the style of inputs and that of the example corpus are different. This paper proposes a method for retrieving "meaning-equivalent sentences" to overcome these two problems. A meaning-equivalent sentence shares the main meaning with an input despite lacking some unimportant information. The translations of meaning-equivalent sentences correspond to "rough translations." The retrieval is based on content words, modality, and tense.

1 Introduction

Speech-to-speech translation (S2ST) technologies consist of speech recognition, machine translation (MT), and speech synthesis (Waibel, 1996; Wahlster, 2000; Yamamoto, 2000). The MT part receives speech texts recognized by a speech recognizer. The nature of speech causes difficulty in translation 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 accurately compared with its performance for written-style text .

Example-based MT (EBMT) is one of the corpus-based machine translation methods. It retrieves examples similar to inputs and adjusts their translations to obtain the output (Nagao, 1981). EBMT is a promising method for S2ST in that it performs robust translation of ungrammatical sentences and requires far less manual work than rule-based MT.

However, there are two problems in applying EBMT to S2ST. One is that the translation accuracy drastically drops as input sentences become long. 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 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 written-style text, which is different from that of natural speech. This style difference makes retrieval of similar sentences difficult and degrades the performance of EBMT.

This paper proposes a method of retrieving sentences whose meaning is equivalent to input sentences to overcome the two problems. A meaning-equivalent sentence means a sentence having the main meaning of an input sentence despite lacking some unimportant information. Such a sentence can be more easily retrieved than a similar sentence, and its translation is useful enough in S2ST. We call this translation strategy example-based "rough translation."

Retrieval of meaning-equivalent sentences is based on content words, modality, and tense. This provides robustness against long inputs and in the differences in style between the input and the example corpus. This advantage distinguishes our method from other translation methods.

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

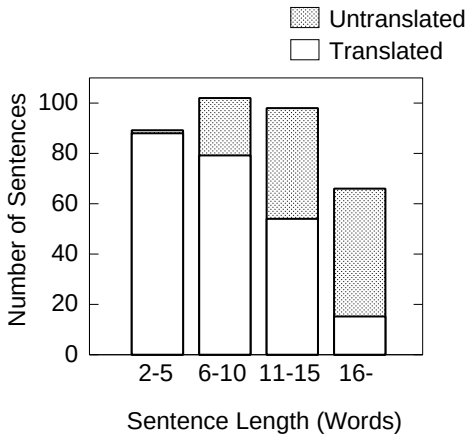

Figure 1: Distribution of Untranslated Inputs by Length

	Language	
	English	Japanese
Concise	5.4	6.2
Conversational	7.9	8.9

Table 1: Number of Words by Sentences

		Language Model	
		Concise	Conversational
Test	Concise	16.4	58.3
	Conversational	72.3	16.3

Table 2: Cross Perplexity

2 Difficulty in Example-based 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 input sentences become, the fewer similar example sentences exist in the example corpus. Figure 1 shows translation difficulty in long sentences in EBMT (Sumita, 2001). The EBMT system is given 591 test sentences and returns translation result as translated/untranslated. Untranslated means that there exists no similar example sentences for the input. Although the EBMT is equipped with a large example corpus (about 170K sentences), it often failed to translate long inputs.

2.2 Style Differences between Concise and Conversational

The performance of example-based S2ST greatly depends on the example corpus. It is advantageous for an example corpus to have a large volume and the same style as the input sentences. A corpus of texts dictated from conversational speech is favorable for S2ST. Unfortunately, it is very difficult to prepare such an example corpus since this task requires laborious work such as speech recording and speech transcription.

Therefore, we cannot avoid using a written-style corpus, such as phrasebooks, to prepare a sufficiently large volume of examples. Contained texts are almost grammatical and rarely contain unnecessary words. We call the style used in such a corpus “concise” and the style seen in conversational speech “conversational.”

Table 1 shows the average numbers of words in concise (Takezawa et al., 2002) and conversational corpora (Takezawa, 1999). Sentences in conversational style are about 2.5 words longer than those in concise style in both English and Japanese. This is because conversational style sentences contain unnecessary words or subordinate clauses, which have the effects of assisting the listener’s comprehension and avoiding the possibility of giving the listener a curt impression.

Table 2 shows cross perplexity between concise and conversational corpora (Takezawa et al., 2002). Perplexity is used as a metric for how well a language model derived from a training set matches a test set (Jurafsky and Martin, 2000). Cross perplexities between concise and conversational corpora are much higher than the self-perplexity of either of the two styles. This result also illustrates the great difference between the two styles.

3 Meaning-equivalent Sentence

Example-based S2ST has the difficulties described in Section 2 when it attempts to translate inputs exactly. Here, we set our translation goal to translating input sentences not exactly but roughly. We assume that a rough translation is useful enough for S2ST, since unimportant information rarely disturbs the progress of dialogs and can be recovered in the following dialog if needed. We call this translation strategy “rough translation.”

We propose “meaning-equivalent sentence” to carry out rough translation. Meaning-equivalent sentences are defined as follows:

meaning-equivalent sentence
(to an input sentence)

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

Important information is subjectively recognized mainly due to one of two reasons: (1) It can be surmised from the general situation, or (2) It does not place a strong restriction on the main information.

	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 2: Examples of Unimportant Information

Figure 2 shows examples of unimportant/important information. Information to be examined is written in bold. The information "*of me*" in (1) and "*around here*" in (3) can be surmised from the general situation, while the information "*of this painting*" in (2) and "*Chinese*" would not be surmised since it denotes a special object. The subordinate sentences in (4) and (5) are regarded as unimportant since they have small significance and are omittable.

3.1 Basic Idea of Retrieval

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

However, functional 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 are given, their relation is almost determined in the domain. Information of modality and tense is extracted from functional words and utilized in classifying the meaning of a sentence (described in Section 3.2.2).

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.

3.2 Features for Retrieval

3.2.1 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 functional 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."

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)

Table 3: Clues for Discriminating Modalities in Japanese

3.2.2 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 tense are distinguished by surface clues, mainly by particles and auxiliary verbs. Table 3 shows a part of the clues used for discriminating modalities in Japanese. Sentences having no clues are classified as others. Figure 3 [2] shows

[1] Number and pronoun are included.

[2] Japanese content words are written in sans serif style and Japanese functional words in *italic* style.

Sentence[3]	Modality & Tense[4]
hoteru *wo* yoyaku *shi* *tekudasai* (Will you reserve this hotel?)	request
hoteru *wo* yoyaku *shi tai* (I want to reserve this hotel.)	desire
hoteru *wo* yoyaku *shi mashi ta ka*? (Did you reserve this hotel?)	question past
hoteru *wo* yoyaku *shi tei masen* (I do not reserve this hotel.)	negation

Figure 3: Sentences and their Modality and Tense

sample sentences and their modality and tense. Clues are underlined.

A speech act is a concept similar to modality in which speakers' intentions are represented. The two studies introduced information of the speech act in their S2ST systems (Wahlster, 2000; Tanaka and Yokoo, 1999). The two studies and our method differ in the effect of speech act information. Their effect of speech act information is so small that it is limited to generating the translation text. Translation texts are refined by selecting proper expressions according to the detected speakers' intention.

3.3 Retrieval and Ranking

Sentences that satisfy the conditions below are recognized as meaning-equivalent sentences.

1. It is required to have 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.

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.

[3]Space characters are inserted into word boundaries in Japanese texts.

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

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

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

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

Figure 4: Example of Conditions for Ranking

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 functional words.
C6: # of different functional words.
(the fewer, the higher priority)

Figure 4 shows an example of conditions for ranking. Content word in a focus area of input are underlined and functional words are written in italic.

4 Experiment

4.1 Test Data

We used a bilingual corpus of travel conversation, which has Japanese sentences and their English translations (Takezawa et al., 2002). 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 example data (Example) and test data (Concise) by extracting test data randomly from the whole set of data.

In addition to this, we used a conversational speech corpus for another set of test data (Takezawa, 1999). This corpus contains dialogs between a traveler and a hotel

[5]Words are converted to base form.

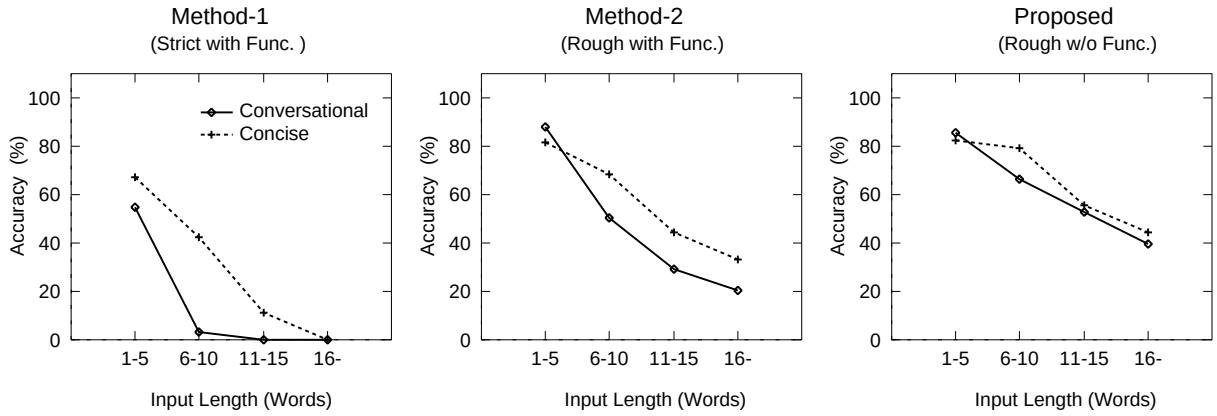

Figure 5: Results

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

Table 4: Statistics of the Corpora

receptionist. It tests the robustness in 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 4.

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

4.2 Compared Retrieval Methods

We use two example-based retrieval methods to show the characteristic of the proposed method. The first method (Method-1) uses "strict" retrieval, which does not allow missing words in input. The method takes functional words into account on retrieval. This method corresponds to the conventional EBMT method. The second method (Method-2) uses "rough" retrieval, which does allow missing words in input, but still takes functional words into account.

4.3 Evaluation Methodology

Evaluation was carried out by judging whether retrieved sentences are meaning-equivalent to inputs. It must be noted that inputs and retrieved sentences are both in Japanese. We did not compare inputs and translations of retrieved sentences, since translation accuracy is a matter of the example corpus and does not concern our method.

The sentence with the highest score among retrieved sentences was taken and evaluated. The sentences are marked manually as meaning-equivalent or not by a Japanese native. A meaning-equivalent sentence includes all important information in the input but may lack some unimportant information.

4.4 Results

Figure 5 shows the accuracy of the three methods with the concise and conversational style data. Accuracy is defined as the ratio of the number of correctly equivalent sentences to that of total inputs. Inputs are classified into four types by their word length.

The performance of Method-1 reflects the narrow coverage and style-dependency of conventional EBMT. The longer input sentences become, the more steeply its performance degrades in both styles. The method can retrieve no similar sentence for inputs 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 inputs. This indicates the robustness of the rough retrieval strategy to longer inputs. 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 inputs 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 Related Work

5.1 EBMT

The rough translation proposed in this paper is a type of EBMT (Sumita, 2001; Veale and Way, 1997; Carl, 1999; Brown, 2000). The basic idea of EBMT is that sentences similar to the inputs 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 functional-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 functional-word part, our method is characterized by disregarding functional words, while other EBMT methods still rely on them for the similarity measure. In our method, the lack of functional 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 to length and the style differences between inputs and the example corpus.

5.2 Translation Memory

Translation memory (TM) is aimed at retrieving informative translation example from example corpus. TM and our method share the retrieval strategy of rough and wide coverage. However, recall is more highly weighted than precision in TM, while recall and precision should be equally considered in our method. To carry out wide coverage retrieval, TM relaxed various conditions on inputs: Preserving only mono-gram and bi-gram on words/characters (Baldwin, 2001; Sato, 1992), removing functional words (Kumano et al., 2002; Wakita et al., 2000), and removing content words (Sumita and Tsutsumi, 1988). In our method, information on functional words is removed and that on modality and tense is introduced instead. Information on word order is also removed while instead we preserve information on whether each word is located in the focus area.

6 Conclusions

In this paper, we introduced the idea of meaning-equivalent sentences for robust example-based S2ST. Meaning-equivalent sentences have the same main meaning as the input despite lacking some unimportant information. Translation of meaning-equivalent sentences corresponds to rough translations, which aim 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.

Then, we described a method for retrieving meaning-equivalent sentences from an example corpus. Retrieval is based on content words, modality, and tense. This strategy is feasible owing to the restricted domains, often adopted in S2ST, which have relatively small variety in lexicon and meaning. An experiment demonstrated the robustness of our method to input length and the style differences between inputs and the example corpus.

Most MT systems aim to achieve exact translation, but unfortunately they often output bad or no translation for long conversational speeches. The rough translation proposed in this paper achieves robustness in translation for such inputs. This method compensates for the shortcomings of conventional MT and makes S2ST technology more practical.

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

T. Baldwin. 2001. Low-cost, high-performance translation retrieval: Dumber is better. In *Proc. of the 39th ACL*, pages 18–25.

R. D. Brown. 2000. Automated generalization of translation examples. In *Proc. of the 18th COLING*.

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

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

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

T. Kumano, I. Goto, H. Tanaka, N. Uratani, and T. Ehara. 2002. A translation aid system by retrieving bilingual news database. In *System and Computers in Japan*, pages 19–29.

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.

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

S. Sato. 1992. CTM: An example-based translation aid system. In *Proc. of the 14th COLING*, pages 1259–1263.

E. Sumita and Y. Tsutsumi. 1988. A translation aid system using flexible text retrieval based on syntax-matching. In *TRL Research Report TR87-1019*. IBM Tokyo Research Laboratory.

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 *Proc. of the 3rd LREC*, 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.

H. Tanaka and A. Yokoo. 1999. An efficient statistical speech act type tagging system for a speech translation systems. In *Proc. of the Association for Computational Linguistics*, pages 381–388.

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

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.

Y. Wakita, K. Matsui, and Y. Sagisaka. 2000. Fine keyword clustering using a thesaurus and example setences for speech translation. In *Proc. of International Conference of Speech Language Processing*, pages 390–393.

S. Yamamoto. 2000. Toward speech communications beyond language barrier - research of spoken language translation technologies at ATR -. In *Proc. of ICSLP*, volume 4, pages 406–411.