Example-based machine translation using DP-matching between word sequences

Eiichiro Sumita
ATR Spoken Language Translation Research Laboratories
2-2 Hikaridai, Seika, Soraku,
Kyoto 619-0288,
Japan
sumita@slt.atr.co.jp

Abstract

We propose a new approach under the example-based machine translation paradigm. First, the proposed approach retrieves the most similar example by carrying out DP-matching of the input sentence and example sentences while measuring the semantic distance of the words. Second, the approach adjusts the gap between the input and the most similar example by using a bilingual dictionary. We show the results of a computational experiment.

1 Introduction

Knowledge acquisition from corpora is viable for machine translation. The background is as follows:

- Demands have been increasing for machine translation systems to handle a wider range of languages and domains.
- MT requires bulk knowledge consisting of rules and dictionaries.
- Building knowledge consumes considerable time and money.
- Bilingual/multilingual translations have become widely available.

There are two approaches in corpus-based translation:

1. Statistical Machine Translation (SMT): SMT learns models for translation from corpora and dictionaries and searches for the best translation according to the models in run-time (Brown et al., 1990; Knight, 1997; Ney et al., 2000).
2. Example-Based Machine Translation (EBMT): EBMT uses the corpus directly. EBMT retrieves the translation examples that are best matched to an input expression and adjusts the examples to obtain the translation (Nagao, 1981; Sadler 1989; Sato and Nagao, 1990; Sumita and Iida, 1991; Kitano, 1993; Furuse et al., 1994; Watanabe and Maruyama, 1994; Cranias et al., 1994; Jones, 1996; Veale and Way, 1997; Carl, 1999, Andriamanankasina et al., 1999; Brown, 2000).

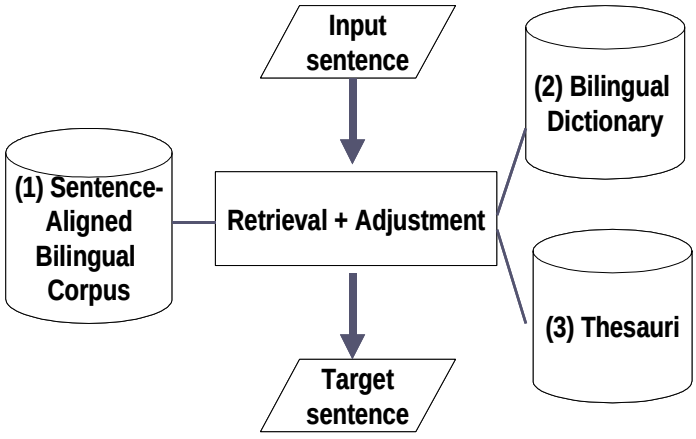

Figure 1 Configuration

This paper pursues EBMT and proposes a new approach by using the distance between word sequences. The following sections show the algorithm, experimental results, and implications and prospects.

2 The proposed method

2.1 Configuration

As shown in **Figure 1**, our resources are (1) a bilingual corpus, in which sentences are aligned beforehand; (2) a bilingual dictionary, which is used for word alignment and translation; and (3) thesauri of both languages, which are used for aiding word alignment and incorporating the semantic distance between words into the word sequence distance.

2.2 Algorithm

The translation process consists of four steps:[1]

I. **Retrieve** the most similar translation pair;
II. **Generate** translation patterns;
III. **Select** the best translation pattern;
IV. **Substitute** target words for source words.

Here, we illustrate the algorithm using translation *from Japanese to English* step by step.

2.2.1 Retrieval - Step I

This step scans the source parts of all example sentences in the bilingual corpus. By measuring the distance (*dist* shown below) between the word sequences of the input and example sentences, it retrieves the examples with the minimum distance, provided the distance is smaller than the given threshold. Otherwise, the whole translation *fails* with no output.

$$(1) dist = \frac{I + D + 2\sum SEMDIST}{L_{input} + L_{example}}$$

$$(2) SEMDIST = \frac{K}{N}$$

According to equation (1), *dist* is calculated as follows: The counts of the Insertion (I), Deletion (D), and Substitution (S) operations are summed up and the total is normalized by the sum of the length of the source and example sequences.

[1] Step I corresponds to *Retrieval* in Figure 1 and steps II, III, and IV correspond to *Adjustment*.

Substitution (S) considers the semantic distance between two substituted words and is called *SEMDIST*. *SEMDIST* is defined as the division of *K* (the level of the least common abstraction in the thesaurus of two words) by *N* (the height of the thesaurus) according to equation (2) (Sumita and Iida, 1991). It ranges from 0 to 1.

Let's observe the following two sentences,[2] (1-j) the input and (2-j) the source sentence of the translation example, where the hatched parts represent the differences between the two sentences.

(1-j) iro/ga/ki/ni/iri/masen
[color/SUB/favor/OBJ/enter/POLITE-NOT]
{I do not care for the color.}

(2-j) dezain/ga/ki/ni/iri/masen
[design/SUB/favor/OBJ/enter/ POLITE-NOT]
{I do not care for the design.}

Because "iro" and "dezain" are completely dissimilar in the thesauri used in the experiment, *SEMDIST* is 1, and therefore, the *dist* between them is (0+0+2*1) / (6+6) = 0.167. The *dist* is calculated efficiently by a standard *dynamic programming* technique (Cormen 1989).

This step is an application of the so-called *DP-matching*, which is often used in speech recognition research.

2.2.2 Pattern Generation - Step II

First, the step stores the hatched parts of the input sentence in memory for the following translation. Second, the step aligns the hatched parts of source sentence (2-j) to corresponding target sentence (2-e) of the translation example by using lexical resources. [3] We do not align non-hatched parts word by word. We assume that non-hatched parts correspond together as a whole. This keeps most parts of the example unchanged in order to avoid mixing errors or unnaturalness in the translation.

(2-j) dezain/ga/ki/ni/iri/masen
(2-e) I do not like the design.

[2] A Japanese sentence has no word boundary marker such as the *blank* character in English so we put « / » between Japanese words. The brackets show the English literal translation word by word and the braces show the sentence translation in English.

[3] We do not consider on the alignment mechanism in this proposal. We have a free hand in selecting an appropriate alignment method out of a spectrum (Manning and Hinrich, 1999) ranging from statistical to lexical types. In the experiment, we rely on a bilingual dictionary and thesauri in both languages.

We obtain the following translation pattern, where the variable X is used to connect source (2-j-p) and target (2-e-p) and store instance (1-j-b) in the input sentence.

(2-j-p) X/ga/ki/ni/iri/masen
(2-e-p) I do not like the X
(1-j-b) X = "iro"

2.2.3 Pattern Selection - Step III

We may retrieve more than one example, and, moreover, translation patterns can differ. We have to select the most suitable one from among these translation patterns. We use a heuristic rule for this purpose.

1. Maximize the frequency of the translation pattern.
2. If this cannot be determined, maximize the sum of the frequency of words in the generated translation patterns.
3. If this cannot be determined, select one randomly as a last resort.

2.2.4 Word Substitution - Step IV

This step is straightforward. By translating the source word of the variable using the bilingual dictionary, and instantiating the variable within the target part of the selected translation pattern by target word (1-e-b), we finally get target sentence (1-e).

(1-e-b) X = "color"
(1-e) I do not like the color.

3 Experiment

To see whether this **rough** approach works or not, we conducted a computational experiment using a large-scale bilingual corpus. In this section, we show the experimental conditions, performance, and error analysis.

Table 1 Corpus Statistics

Sentences	204,108
Sentence Length	(J) 8.3
	(E) 6.1
Words	(J) 1,689,449
	(E) 1,235,747
Vocabulary	(J) 19,640
	(E) 15,374

3.1 Experimental Conditions

Bilingual Corpus
We built a collection of Japanese sentences and their English translations, which are usually found in phrasebooks for foreign tourists. Because the translations were made sentence by sentence, the corpus was sentence-aligned by birth. We *lemmatized and POS-tagged* both the Japanese and English sentences using our morphological analysis programs. The total sentence count was about 200 K. [4] The statistics are summarized in **Table 1**.

Test set
A quality evaluation was done for 500 sentences selected randomly from the above-mentioned corpus and the remaining sentences were used as translation examples for the experiment.

Bilingual Dictionary
We also used a bilingual dictionary previously developed for another MT system in the travel domain (Sumita et al.1999).

Thesaurus
We used thesauri whose hierarchies are based on the Kadokawa Ruigo-shin-jiten (Ohno 1984) for distance calculation and word alignment.

3.2 Results

Here, we show coverage and accuracy results as evidence that our proposed machine translation system works.

3.2.1 Coverage

Our approach does not produce any translation when there is no example whose *dist* is within the given threshold, which was 1/3 in the experiment.

Table 2 Coverage and Sentence Length

	%	*Average length*
Exactly	46.4	5.6
Approximately	42.8	7.7
No Output	10.8	11.0
Total	100.0	7.0

Our approach covers about 90% of 500 randomly selected sentences. As shown in **Table 2**, one half of 90% is matched *exactly* and the other half is matched *approximately (dist < 1/3).*

[4] We call a sequence of sentences uttered by a single speaker an utterance. Our corpus is in fact aligned utterance by utterance. Strictly speaking, 'sentence' in this paper should be replaced by 'utterance.'

The characteristics of *no output* sentences are clearly explained by the average length. Our approach is not good with longer sentences because our algorithm has no explicit step of decomposing an input sentence into sub-sentences and because the longer the sentence, the smaller the possibility that there exists a similar sentence in the example database.

We assume that a coverage of 90% is important because this means that if 200 K sentences were input into the system, the system would produce a translation 90% of the time. In other words, the system would help the user 90% of the time to communicate with foreign people (assuming the user to be in a foreign country).

3.2.2 Accuracy

Quality Ranking

Each translation is graded into one of four ranks[5] (described below) by a bilingual human translator who is a native speaker of the target language, American English:

(A) Perfect: no problems in either information or grammar; (B) Fair: easy-to-understand with some unimportant information missing or flawed grammar; (C) Acceptable: broken but understandable with effort; (D) Nonsense: important information has been translated incorrectly.

Table 3 Translation Accuracy

	Rank	*%*
Good	A	41.4
	B	25.2
	C	11.8
Bad	D	10.8
	F(*No output*)	10.8

Result

As shown in **Table 3**, our proposal achieved a high accuracy of about 80% (A, B, C ranks in total). The remaining 20% are divided into ranks D and F (*No output*).

Long Sentence Problem

Figure 2 shows[6] that **the accuracy clearly decreases as the *dist* increases**. This implies two points: (1) *dist* can indicate the quality of the produced translation, in contrast with the fact that MT systems usually do not provide any confidence factor on their results. The user is safe if he/she confines himself/herself to using translations with a small *dist* value; (2) The current algorithm has a problem in handling distant examples, which usually relate to the *long sentence* problem.

3.2.3 Error Analysis

As shown in the previous two subsections, the most dominant problem is *in dealing with relatively longer sentences*. We point out here that even for shorter sentences there are problems, although they are less frequent, as follows:

- **Idioms or collocations**

Even when the *dist* between the two sentences is small, i.e., they are quite similar in the source language, the meanings of the sentences can vary and the translation can be different in the target language. This case is not so frequent, but is possible by idioms or collocations as exemplified in the following sample.

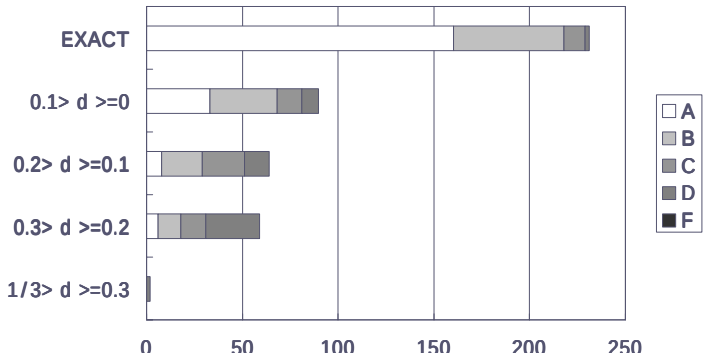

Figure 2 Accuracy by Distance

[5] This ranking was developed for evaluation in spoken language translation. For more details, see (Sumita et al., 1999).

[6] The horizontal axis indicates the number of sentences.

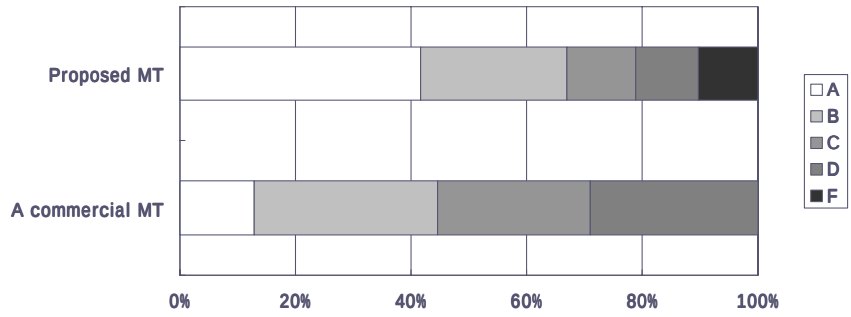

Figure 3 Comparison of Proposed EBMT and a Commercial MT

1. kata/o/tsume/teitadake/masu/ka [shoulders/OBJ/shorten-or-sit-closely/REQUEST/POLITE/QUESTION] {Could you tighten the shoulders up?}
2. seki/o/tsume/teitadake/masu/ka [seat/OBJ/shorten-or-close-up/REQUEST/POLITE/QUESTION] {Could you move over a little?}

The replaceability between "kata" and "seki" does not hold for these two similar sentences. To avoid this problem, a feedback mechanism of erroneous translations built into the system is one possible solution.

- **Noise in data**

The proposed approach accepts the translation example blindly. If the translation is *wrong* or *inappropriate*, the output is directly made defective. The next two translations show contextual inappropriateness. The source parts of the two examples are exactly the same, but the target part of the first example is neutral and that of the second example is specific, i.e., valid only in a special situation. Preventing this requires cleaning the example database, preferably by machine, or collecting sufficiently large-scale data to suppress the influence of noisy examples.

1. hai/ari/masu = Yes, we do. [yes/exist/POLITE]
2. hai/ari/masu = Yes, we have a shuttle bus. [yes/exist/POLITE]

4 Discussion

Here, we explain the implications of the experimental results and discuss the future extension.

4.1 Limitations

Our proposal has the limitations listed below, but we would like to note that we have obtained high coverage and accuracy for the phrasebook task.

- Database limitation: If a *nearest neighbor* within the threshold does not exist in the example database, we cannot perform translation. One positive note is that we were able to build the necessary example database for the phrasebook task, which is not a toy.
- Context limitation: We cannot translate context-dependent words, because contexts are often hard to embed in an example database. For example, Japanese 'konnichiwa' corresponds to 'Good morning' or 'Good afternoon' in English depending on the time of utterance. It is in general difficult to embed such kinds of situational information into the example database.
- Implementation limitation: We have no method for dividing an input into chunks (such as clauses) at present, so long sentences cannot be dealt with. In addition, no investigation has been made on robustness with respect to recognition errors yet. However, DP-matching is expected to be effective.

4.2 Generality vs. Quality

There is no commercial system that can translate phrasebook sentences at this level of accuracy. **Figure 3** shows a comparison of our proposal and a commercial machine translation system that *accepts* travel conversations. **Table 4** shows sample translations by the above two systems. The upper translation was produced by our proposed system and the lower translation was produced by the commercial system.

The reason behind the performance difference for this task is that the commercial one was built as a general-purpose system and phrasebook sentences are not easy to translate into high-quality results by using such a general-purpose architecture.

However, we must admit that general-purpose architectures are effective and we do not mean to

Table 4 Sample translations by two MTs

pan/o/motto/itadake/masu/ka [bread/OBJ/more/get/POLITE/QUESTION]	Could I have some more bread?
	Is it possible to have bread more?
suteeki/no/yaki/kagen/wa [steak/of/grill/degree/TOPIC]	How would like your steak?
	The roasting addition and subtraction of the steak.
Sentaku/ki/no/tsukai/kata/o/oshie/tekure/masen/ka [washing/machine/of/use/way/OBJ/teach/REQUEST/POLITE/NEGATION/QUESTION]	Will you show me how to use the washing machine?
	Don't I let me know a way of using a washing machine.
eigo/wa/tokui/dehaari/masen [English/TOPIC/strong/be/ POLITE/NEGATION]	I'm not good at English.
	English isn't good.
korera/no/en/wo/doru/ni/ryougae/shi/tai/n/desu/ga [these/of/yen/OBJ/dollar/IND-OBJ/exchange/do/want/PARTICLE/be/but]	I'd like to change yen into dollars.
	But made to want to change these yen into dollars.

criticize them by using this comparison. It is reasonable to suggest that general-purpose architectures are not the most suitable option for achieving high-quality translations in restricted domains.

4.3 Development Cost and Its Reduction in the Future

We do not need grammars and transfer rules but we do need a bilingual corpus and lexical resources.

The adaptability of our approach to multilingual translation is promising. Because we have already succeeded in J-to-E, one of the most difficult translation pairs, we have little concern about other pairs. If we can create an *n* lingual corpus, we can make *n(n-1)* MT systems.

To enable such a dream within a shorter timeframe, we have to reduce the necessary resources such as bilingual dictionaries and thesauri by automating the construction of lexical knowledge.

We are aiming at such additional cost reduction.

We also want to eliminate restrictions, e.g., sentence-aligned and morphologically tagged example database. By doing so, the applicability of our approach can be increased. This is another important challenge.

A further challenging goal is to establish technology enabling the use of a small-scale corpus.

4.4 Related Research

Here, we would like to compare our proposal and related research in four points: level of knowledge, application of dynamic programming, the use of thesauri, and the task.

Knowledge of EBMT

Many EBMT studies (Sato and Nagao, 1990; Sato, 1991; Furuse et al., 1994; Sadler, 1989) assume the existence of a bank of aligned bilingual *trees* or a set of *translation patterns*. However, building such knowledge is done by humans and is very expensive. Methods for automating knowledge building are still being developed. In contrast, our proposal does not rely on such a high-level analysis of the corpus and requires only word-level knowledge, i.e., morphological tags and dictionaries.

Dynamic programming

Dynamic programming has been used within the EBMT paradigm (1) for technical term translation (Sato, 1993), and (2) for translation support (Cranias et al., 1994).

Sato translates technical terms, which are usually compound nouns, while we translate sentences. He uses a corpus in which translation units of a pair of technical terms are aligned, while we do not require the alignment of translation units. He defines the matching score and we define the distance between word sequences, which are different. However, both are computed by a standard dynamic programming technique.

Based on surface structures and content words, Cranias defined a similarity score between texts and introduced the idea of clustering the translation memory to speed up the retrieval of similar translation examples. The score is again computed by a standard dynamic programming technique, but Cranias provides not a translation but only a retrieval.

Thesaurus

Brown (2000) uses *equivalence classes* to successfully improve the coverage of EBMT. He proposed a method of automatically generating equivalence classes using clustering techniques, while we use hand-coded thesauri (in the experiment). Such automation is very attractive, and the author is planning to follow in Brown's line, in spite of a fear that low frequent words will not be dealt with effectively by clustering techniques. Brown uses a hard condition, i.e., whether a word is included in an equivalence class or not, while we provide the relative distance between two words. It is unknown which method is better for EBMT. We do not plan on sticking with the current implementation using hand-coded thesauri, as we realize that further research on these open problems is indispensable.

Phrasebook task

The phrasebook task was first advocated for the task of speech translation by (Stentiford and Steer, 1987). They pointed out that when communicating within a limited domain such as *international telephone communications*, it is nearly possible to specify all of the required message concepts. They used a *keyword-based approach* to access concepts to overcome speech recognition errors. On the other hand, we use *DP-matching techniques* for this end.[7] The scalability of the keyword-based approach has raised questions because enlarging a corpus directly increases the chances of conflict in identifying the concepts to be conveyed.

5 Concluding Remarks

We proposed a new approach using DP-matching for retrieving examples within EBMT and demonstrated its coverage and accuracy through a computational experiment for a restricted domain, i.e., a phrasebook task for foreign tourists.

There is much room for our translation method to improve: (1) decomposing input sentences will improve the coverage, and (2) indexing or clustering the example database will drastically improve the efficiency of the current naïve implementation.

Acknowledgement

First, the author thanks anonymous reviewers for their useful comments. The author's heartfelt thanks go to Kadokawa-Shoten for providing the Ruigo-Shin-Jiten. Thanks also go to Dr. Seiichi YAMAMOTO, President and Mr. Satoshi SHIRAI, Department Head, for providing the author with the chance to pursue this research.

[7] We believe our approach is robust against speech recognition errors, but we have not yet applied it to speech recognition results.

References

Andriamanankasina, T., Araki, K. and Tochinai, T. 1999. *Example-Based Machine Translation of Part-Of-Speech Tagged Sentences by Recursive Division*. Proceedings of MT SUMMIT VII. Singapore.

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J., Mercer, R. L., and Roossin, P. S. 1990. *A Statistical Approach to Machine Translation*. Computational Linguistics 16(2).

Brown, R. D. 2000. *Automated Generalization of Translation Examples*. In Proceedings of the Eighteenth International Conference on Computational Linguistics (COLING-2000), pp. 125-131. Saarbrücken, Germany, August 2000.

Carl, M. 1999. *Inducing Translation Templates for Example-Based Machine Translation*, Proc. Of MT-Summit VII.

Cormen, H. T., Leiserson, C. E. and Rivest, L. R. 1989. *Introduction to Algorithms*, MIT Press, p. 1028.

Cranias, L., Papageorgiou, H. and Piperidis, S. 1994. *A Matching Technique in Example-Based Machine Translation*. Institute for Language and Speech Processing, Greece. Paper presented to Computation and Language.

Furuse, O., Sumita, E. and Iida, H. 1994. *Transfer-Driven Machine Translation Utilizing Empirical Knowledge*. Transactions of IPSJ, Vol. 35, No. 3, pp. 414-425 (in Japanese).

Jones, D. 1996. *Analogical Natural Language Processing*. UCL Press. London, 155p.

Kitano, H. 1993. *A Comprehensive and Practical Model of Memory-Based Machine Translation*. Proc. of IJCAI-93. pp. 1276-1282.

Knight, K. 1997. *Automating Knowledge Acquisition for Machine Translation*, AI Magazine, 18/4.

Manning, D., C. and Hinrich, S. (1999) Chapter 5 of *Foundations of statistical natural language processing*, MIT Press, p. 680.

Nagao, M. 1981. *A Framework of a Mechanical Translation between Japanese and English by Analogy Principle*, in Artificial and Human Intelligence, A. Elithorn and R. Banerji (eds.) North-Holland, pp. 173-180, 1984.

Ney, H., Och, F. J. and Vogel, S. 2000. *Statistical Translation of Spoken Dialogues in the Vermobil System*, Proc. Of MSC2000, pp. 69-74.

Ohno, S. and Hamanishi, M. 1984. *Ruigo-Shin-Jiten*, Kadokawa, p. 932 (in Japanese).

Sadler, V. *Working with Analogical Semantics*, 1989). Foris Publications, p. 256.

Sato, S. and Nagao, M. 1990. *Toward Memory-based Translation*. In the proceedings of the International Conference on Computational Linguistics, COLING-90, Helsinki, Finland, August 1990.

Sato, S. 1991. *MBT2: A Method for Combining Fragments of Examples in Example-Based Translation*. JJSAI, Vol. 6, No. 6, pp. 861-871 (in Japanese).

Sato, S. 1993. *Example-Based Translation of Technical Terms*. Proc. of TMI-93, pp. 58-68,.

Stentiford, F. M. W. and Steer, M. G. 1987. *A Speech Driven Language Translation System*, Proc. of European Conference on Speech Technology, Vol. 2, pp. 418-421.

Sumita, E. and Iida, H. 1991. *Experiments and Prospects of Example-Based Machine Translation*. Proc. of ACL-91, pp. 185-192.

Sumita, E., Yamada, S., Yamamoto, K., Paul, M., Kashioka, H., Ishikawa, K. and Shirai, S. 1999. *Solutions to Problems Inherent in Spoken-language Translation: The ATR-MATRIX Approach*, Proc. of 7th MT Summit, pp. 229-235.

Veale, T. and Way, A. 1997. *Gaijin: A Template-Driven Bootstrapping Approach to Example-Based Machine Translation*, in the Proceedings of NeMNLP'97, New Methods in Natural Language Processing, Sofia, Bulgaria.

Watanabe, H. and Maruyama, H. 1994. *A Transfer System Using Example-Based Approach*. IEICE Transactions on Information and Systems, Vol. E77-D, No. 2, pp. 247-257.