自然言語処理　120－3
（1997.　7.　24）

実例に基づく機械翻訳手法における類似文検索手法

アンドリアマナカシナ タンテリ† 荒木 健治‡ 宮永 喜一† 栃内 香次†

†北海道大学大学院工学研究科
〒０６０札幌市北区北１３条西８丁目
電話 011-706-6823 FAX 011-709-6277
e-mail : {tantely,miyanaga,tochinai}@hudk.hokudai.ac.jp

‡北海学園大学工学部
〒０６４札幌市中央区南２６西１１
電話 011-811-1161 FAX 011-551-2951
e-mail : araki@eli.hokkai-s-u.ac.jp

あらまし　本稿文は実例に基づく機械翻訳における入力文と類似文した実例を検索するマッチング手法について述べる。品詞タグの分析によって入力文と実例文の間で、類似している文を探索する。本手法は隣接する語が深い関係を持っていることに基づいている。それゆえ、実例文の単語と入力文単語に関係が高ければ、それぞれに隣接する単語関の関係も高くなると考えられる。本手法は仏日翻訳システムに応用されている。実験は仏日の会話についての文章本を用いて行った。

キーワード　実例に基づく翻訳、機械翻訳、マッチング手法、翻訳支援，話し言葉の翻訳

Method for Searching the Best-Matching Sentence in Example-Based Machine Translation

Tantely Andriamanankasina† Kenji Araki‡ Yoshikazu Miyanaga† Koji Tochinai†

†Faculty of Engineering, Hokkaido University
N-13 W-8 Kita-ku, 060 Sapporo
Tel 011-706-6823 Fax 011-709-6277
e-mail : {tantely,miyanaga,tochinai}@hudk.hokudai.ac.jp

‡Faculty of Engineering, Hokkai-Gakuen University
S-26 W-11 Chuo-ku, 064 Sapporo
Tel 011-811-1161 Fax 011-551-2951
e-mail : araki@eli.hokkai-u-s.ac.jp

Abstract　This paper proposes a matching method for the Example-Based Machine Translation. The method attempts to find a stronger match between the input sentence and the source sentence from the translation examples archive, by a profound analysis of the part-of-speech tags. The idea is based on the fact that words have a close relationship with their direct neighbors in a sentence. Therefore, the relationship between two words is much stronger if it is supported by its direct neighbors' correspondence. The method is implemented in a French-Japanese translation system, and experiments are done with spoken-language text taken from French-Japanese conversation books.

key words　example-based translation, machine translation, matching method, translation aids, spoken-language translation

1. Introduction

The idea of Example-Based Machine Translation (EBMT) was first proposed by Nagao [Nagao 84]. It is essentially translation by analogy : having a bilingual translation examples archive, a given input sentence is compared with the source-language side of the archive. The system selects the most similar match and performs a transfer operation.

The final goal of our research is to build up a French-Japanese translation system based on the example-based method. Various models have been proposed [Sato and Nagao 90], [Kitano 93] and different issues are discussed [Jones 96]. We present in this paper some improvements in the matching method. During the transfer, we try also to build up language-dependent rules to handle missing words or surplus of word efficiently. Missing word (resp. surplus of word) is words which is present (resp. missing) in the input sentence but missing (resp. present) in the source sentence of the example translation. As for the sentence unit, we do not tackle the boundary friction and incorrect chunking problem. We assume that the text unit is a whole sentence.

The most popular similarity metric is the word-based metric. It compares words of the two sentences in terms of their morphological paradigm, hypernonyms, hyponyms, antonyms, and part-of-speech (pos) tags [Niremburg 93]. Others are character-based metric, [Sato 92], which considers some characteristics of the Japanese language. Syntax-rule driven metric seems very promising because it tries to capture similarity at the syntax level. Cranias [Cranias 94] tried to capture this kind of similarity through the observation of pos tags and functional words data. The method proposed here searches for a stronger correspondence between the input sentence and the retrieved example. Words are ambiguous and can take on different functions in different sentences. The nature of their direct neighbors, in a sentence, can help to determine exactly what they are. During the search for the best match, and the calculation of the similarity metric, we observe simultaneously two consecutive words instead of only a single word. We propose also a simpler but better method which searches for the best match, based only on pos tags.

In the next section, the overview of the whole translation is presented, with the structure of a translation example. The presentation of the part-of-speech tags list is given before the matching method. Finally, experiment and results are discussed and few words are given as conclusion.

2. Overview of the translation

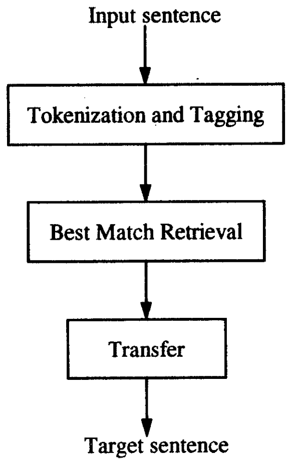

Figure 1 : Overview of the whole translation

The translation system is divided into 3 steps :
First, input sentence is tokenized and tagged. The tokenization program works as follows : blanks are first put between any punctuation, or a hyphen, and after an apostrophe. After this, exceptions are assembled again,. The list of exceptions is very long, it is almost composed of words whose combination forms one tag. For example, *un_peu_plus*, *-t-il*, *tout-de-suite*, *y_compris*, *New-York*. After this, a tokenized input sentence is tagged, using a language lexicon. Disambiguation is not performed and

all possible tags are taken. We use a sub-lexicon from the lexicon developed at the INaLF (Institut National de Langue Française) [INaLF 97].

Secondly, matching operation is performed, where 5 best matches are selected. However, this matching method will be explained in section 4. During the same time that the matching operation is performed, the program collects translations examples whose source sentence contains words of input sentences. Its results are saved in a table and indexed by word position in the input sentence. Each entry contains the 3 best translation examples found. These are those containing the longest series of exact word matches, starting from the given index. During the transfer operation, these translations will be used as a dictionary, to replace words in the translation example, which differ from the input sentence.

At last, the transfer is performed by considering some characteristics of both languages. When a word is missing, some words which depend on that word must also be deleted. For example, in the Japanese language, there are the particles は or を which do not have correspondence in French. We, therefore, try to construct some language-dependent rules, like :

1. In the Japanese language, particles depend on the word preceding them.
2. In the French language, determinants depend on the noun following them.

A translation example [Table 1] is composed by the French source sentence, its Japanese translation sentence, and a map describing the correspondence between words in both sentences. Every word in the French source sentence is followed by its part-of-speech tag. As for the Japanese, the structure *surface pronunciation semantic tags* from the result of the Japanese tagging program is used, for simplicity. The correspondence map is of the form *WPF1,WPF2,.../WPJ1,WPJ2,...*, where *WPFi* is word position in the French sentence, and *WPJj* is word positions in the Japanese sentence.

Table 1 : A translation example

French sentence : *elles/PRV sont/ECJ au/DTC premier/ADJ etage/SBC ./.*

Japanese sentence :

二	に	二	数詞		
階	かい	階	普通名詞		
に	に	に	格助詞		
あり	あり	あり	動詞	子音動詞ラ行	基本連用形
ます	ます	ます	動詞性接尾辞	動詞性接尾辞ます形	基本形
。	。	。	句点		

EOS

Correspondence maps : *2/4 3/3 4,5/1,2*

3. Description of the part-of-speech tags list

Since the result of the matching depends mostly on the tag list, it is worth being presented here. The more the tag lists are detailed, the better match can be obtained, if it appears in the translation examples. However, to make the system able to work with restricted translation examples, as usually is the case, restrictions should be taken.

The tag list used, came largely from the tag list proposed by the InaLF for the French language. They proposed 70 pos tags in which 20 are punctuations and 3 are for particular cases. We removed differences between plural and singular, because they make a negative contribution to capture similarity of the two sentences. The system uses therefore 48 pos tags. They are characterized as follows :

1. Verbs *être* (to be) and *avoir* (to have) are separated from the rest of the verbs. This is justified by the special use of these verbs in the language. Any verb may take 4 tags depending on its form : infinitive, conjugated, past participle, or *-ant* (-ing) form.

2. Combined determinant, like *des*, *du*, and simple determinant are treated as belonging to different groups.
3. Pronouns are divided in two groups : those supported by the verb, and the others.
4. Simple adjectives are separated from past participle form adjectives, which may take different pos tags depending on whether they follow the verb *être* or not.
5. Among the rest of the tags, there are nouns, adverb, cardinal, preposition, relative.

4. Matching method and Similarity Metric

The method searches for the shortest sentence having the largest number of consecutive word matches. Word comparison is based on pos tags. The search works as follows : every sentence in the translation examples is compared with the input sentence. Words are compared one by one. Only equivalence supported by an equivalence between one of the words' direct neighbor is considered. That means : if a word at a position i in the sentence A has the same tag as the word at the position j in the sentence B, the word at the position $i-1$ or $i+1$ in A must have a same tag as the word at the position $j-1$ or $j+1$. A virtual word is put at the beginning of each sentence to make this test valid for any length word sentence, especially one-word sentences. Since disambiguation is not performed during the tagging operation, a word $W1$ of the input sentence is considered to be equivalent to word $W2$ of the source sentence, if the pos tag of $W2$ appears in the pos tag list of $W1$. Result of this operation is saved in a list ordered by WPI + WPS, where WPI is the word position in input sentence and WPS, is the word position in the source sentence.

Deletion operation performed on this list starts from left to right with doubles and crossings being eliminated. Doubles appear when one word has many correspondences in the other sentence. Crossing indicates equivalence, breaking the normal sequence (left to right) : Assuming that word $I1$ has a correspondence $S1$, and word $I2$, $S2$, crossing appears if $WPI1 < WPI2$ but $WPS1 > WPS2$.

To find the shortest sentence which have the largest number of word matches, the following similarity metric is proposed :

$$M = 100*NT - 10*L + NW,$$

where, NT is the number of consecutive words matches found, L is the length of the source sentence, and, NW is the number of words exactly match. The bigger this value is, the closer the two sentences are. The method is illustrated in Table 2. The deletion process does not modify anything in this case, since there is no double or crossing.

Table 2 : Calculation of the Similarity Metric

Input Sentence : *ou/REL est/ECJ -ce/PRV que/SUB,SUB$,PRO je/PRV pourrais/VCJ trouver/VNCFF des/DTC cctelettes/SBC de/PREP porc/SBC ?/?*	
Source Sentence : *quand/SUB est/ECJ -ce/PRV que/SUB$ vous/PRV passerez/VCJ le/DTN concours/SBC d'/PREP admission/SBC ?/?*	11 words
Correspondences found : *est -ce/est -ce, ce que/ce que, que je/que vous, vous pourrais/vous passerez, cotelettes de/concours d', de porc/d' admission, porc ?/admission ?,*	7 correspon dences
Exact Matches : *est, -ce, que, ?.*	4 matches
Similarity Metric $M = 7*100 - 11*10 + 4 = 604$	

5. Experiments and results

The translation examples contains 1098 French-Japanese pairs, taken from French-Japanese Conversation books [Meguro 87] and [Sato F. 90]. One French sentence has an average of 8 words. It is justified by the presence of a short sentence in the spoken-language world, like *Quoi ?*

(What ?). We train the French source sentences of the translation examples, with the tagging program developed by the INaLF, and the Japanese sentence with the Japanese language tagging program [Nara-aist 97]. Results are corrected manually.

Building up the sub-sentential alignments was the hardest work, since the ideas proposed so far [Gale 91] [Fung 94] assume the availability of a very large corpora.

The system is tested with 367 French input sentences, taken arbitrarily from the same source. We have performed two experiments : one concerns the global result of the matching method and the other concerns a comparison of the method with the case where only a single word comparison is considered instead of consecutive words.

In the first experiment, we observed the structure of the Japanese sentence, which is translation of the first best match. We divided them into 3 categories :

1. The proposal has (or almost has) the same structure as the translation. It is characterized by the presence of all elements of the target sentence. Production of the translation is almost an operation of replace.
2. The proposal can be used in order to produce the translation. This is represented by sentences where some surplus or missing words exist, but the structure of the sentence is the same.
3. The proposal does not help to produce the translation. Those are sentences which have completely different structure.

The result is presented in Table 3. It shows that 67% of the whole proposals can be used to produce the translation.

Table 3 : Qualities of the translation of the best match

Categories	*Results*
1	39 %
2	28 %
3	33 %

As for the second experiment, we compare the method with a system based on single word comparison. Words are compared without considering what their direct neighbors are. In the similarity metric, the number of single word matches is counted in the place of the number of consecutive words matches. We observed only those proposals which differ from the proposals of the method, and divided them into 4 categories :

1. Both proposals can be judged as having the same structure and can be used to produce the translation.
2. Both proposals are almost the same and do not help to produce the translation.
3. The proposal of the method is better than the result of the single word-based method.
4. The proposal of the method is worst than the result of the single word-based method.

The result is presented in Table 2. We found 120 differences. The method lost 10/367 from the single word-based method, but won 38/367. This means a net improvement of 28/367, or 8% of the whole system.

Table 2 : Comparison with single word based method

Categories	*Results*
1	22
2	50
3	38
4	10
Total	120

6. Discussion

The result shows that this method improves the single word-based method. Before making a final judgment on this issue, we intend to conduct more extensive experiments, with larger translation examples and more input sentences and evaluate the result at the final output of the translation. The 33 % of fails in the Table 3 are almost due to lack of examples. However, since the method observes only pos tags, some changes of sentence structure because of word meaning are also recorded. The 8 % improvement in Table 4 is very promising. The 10 fails recorded in Table 4, category 4 shows that the number of consecutive word matches in the sentence is too small to capture the structure of the sentence, isolated word plays a very important role in these success of the single word-based method. Re-discussion of the tag list would solve these problems, however, working with larger translation examples is preferred.

7. Conclusion

In this paper, we have described and tested a matching method, which searches a stronger relationship between two sentences. In the method, correspondence between 2 words is supported by correspondences between their direct neighbors. It Is also characterized by its simplicity : only profound observation on pos tags is performed.

Preliminary results are very promising. A test with larger data has yet to be performed and the results at the level of the whole translation (real output sentences) have to be observed.

8. References

[Cranias 94] L. Cranias, H. Papageorgiou, and S. Piperidis. 1994. A Matching Technique in Example-Based Machine Translation. *Institute for Language and Speech Processing, Greece.* Paper presented to *Computation and Language http://xxx.yukawa.kyoto-u.ac.jp/archive/cmp-lg*

[Fung 94] P. Fung and K.W. Church. 1994. K-vec : A New Approach for Aligning Parallel Texts. *In COLING-94, Kyoto, Japan,* pp 1096-1101

[Gale 91] W.A. Gale and K.W Church. 1991. A Program for Aligning Sentences in Bilingual Corpora. *Proceedings. of the 29th Annual Meeting of the ACL.,* pp 177-184

[INaLF 97] Demonstration du catégorisateur d'Eric Brill (version 1.14) entrainé pour le français à l' INaLF *http://sun1.inalf.ciril.fr/brillinalf/ebt.pres.html*

[Jones 96] D. Jones. 1996. Analogical Natural Language Processing. *UCL Press. London.*

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

[Meguro 87] S. Meguro. 1987. Manuel de conversation française. *Hakusuisha. Tokyo.*

[Nagao 84] M. Nagao. 1984 A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. *Artificial an' Human Intelligence, ed. Elithorn A. and Banerji R., North-Holland,* pp 173-180.

[Nara-aist 97] Nara Institute of Science and Technology. 1997. Chasen1.0 home page. *http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html.*

[Niremburg 93] S. Niremburg. et al. 1993. Two approaches to Matching in Example-Based Machine Translation. *Proceedings of TMI-93, Kyoto, Japan.*

[Sato 92] S. Sato. 1992 CTM : An Example-Based Translation Aid System. *Proceedings. of COLING-92* pp 1259-1263.

[Sato and Nagao 90] S. Sato and M. Nagao. 1990. Toward Memory-based Machine Translation. *In Proceedings of COLING-90,* pp 247-252.

[Sato F. 90] F. Sato. 1990. Locutions de base. *Hakusuisha. Tokyo.*