SUB-SENTENTIAL ALIGNMENT METHOD BY ANALOGY

Tantely Andriamanankasina, Kenji Araki, and Koji Tochinai

Graduate School of Engineering,
Hokkaido University, Japan
Email: {tantely,araki,tochinai}@media.eng.hokudai.ac.jp

ABSTRACT

This paper describes a method for searching word correspondences between pairs of translation sentences. In the Example-Based Machine Translation, translation patterns can be extracted easily if word correspondences between pair of translation sentences are defined. The popular methods for aligning bilingual corpus at a sub-sentential level are unable to produce reliable result when the size of the corpus is limited, because they are based on statistics. We propose a method for incrementing a word correspondence-included initial corpus automatically. It is appropriate for new languages whose huge corpus as well as machine readable dictionaries are still not available. The method was evaluated with French-Japanese spoken language texts. As the number of translation examples goes beyond 1,000, more than 80.0% of correct word correspondence rates were earned.

1. INTRODUCTION

Being able to define accurately and efficiently word correspondences between parallel texts is an issue in corpus-based machine translation. Translation patterns can be extracted very easily if word correspondences between pairs of translation sentences are defined. For sub-sentential alignment, a simple consultation of a machine readable dictionary seems to be a very obvious method. However, the presence of unregistered words, and the difference which may be seen in the surface form of words, or after being processed by a lexical analyzer cause problems. In addition, compound words, which often appear in one-to-many or many-to-many word correspondences, are not covered by single word entry-based dictionaries. The use, not of dictionaries, but of the parallel corpus itself, has been therefore suggested in order to search word correspondences [4, 5, 6].

Sub-sentential alignment methods which have been proposed are based on statistics. The problem with statistical methods is that they are not able to produce reliable results when the corpus size is limited. For unexplored languages whose huge parallel corpus are still not available, these approaches cannot be applied. In addition, correspondences involving multiple tokens have not yet been entirely resolved [4]. Melamed avoids these cases and considers only one-to-one correspondence [5]. On the other hand, although Kitamura' s [6] main goal is not the sub-sentential alignment, but the extraction of translation patterns, possible correspondences between functional words are not considered. It weakens the method since it decreases the number of translation patterns which could be extracted.

We have proposed a French-Japanese example-based machine translation which uses word correspondence-included translation examples [1, 2]. As a relatively new field, a huge parallel corpus is not available. Manual construction of translation examples are very time consuming. An automatic method which can efficiently be applied to size-limited corpus is necessary.

In this paper, we propose a method for estimating the word correspondences between new pairs of translation sentences by analogy. Similar pairs are selected from a word correspondences-included initial translation examples. Word correspondences between the new pairs are predicted according to the word correspondences between these similar pairs. By doing so, the method can work with size-limited corpus. Besides, it is considered to be able to resolve efficiently correspondences involving multiple tokens, since solution models are given beforehand within the translation examples. For example, if **"voulez - vous"**

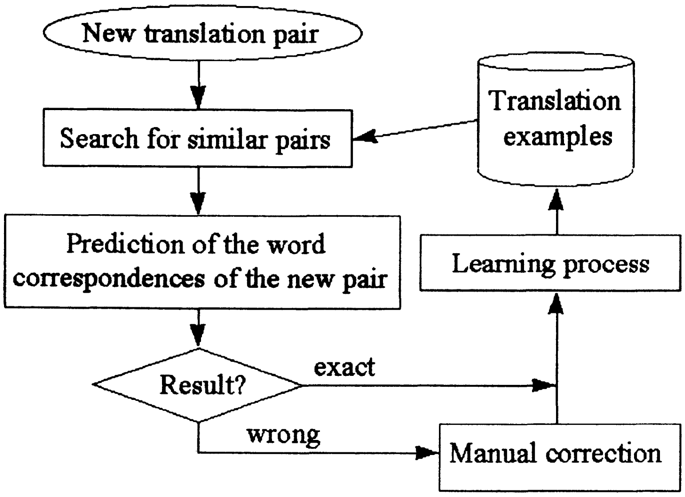

Figure 1: Overview of the method

Table 1: Structure of a translation example

French Sentence	**vous/PRV avez/ACJ un/DTN cendrier/SBC ?/?**
Japanese Sentence	*haizara*/6 *wa*/9 *ari*/2 *masu*/14 *ka*/9 ./1
Correspondence Map	2/3 4/1 5/6

PRV: pronoun, DTN: determinant, SBC: common noun, ACJ: verb "avoir"
1: punctuation, 2: verb, 6: noun, 9: particle, 14: suffix

(would you) and *"itadake masu ka"* appear in a pair of translation sentences, aligning **"voulez"** with *"itadake"*, or **"voulez - vous"** with *"itadake"*, or else **"voulez - vous"** with *"itadake masu ka"* are all conceivable. Observation of sentence meaning sometimes produces results which differ from statistic results. Besides, results depend on whether only one-to-one correspondences are considered or not. However, if there is a translation example mapping **"voulez - vous"** or its similar phrase **"pouvez - vous"** (could you) with *"itadake masu ka"*, that map can be used as a reference to determine the correspondence maps of the new pair.

We introduce, in addition, a learning process-based method for constructing gradually the translation examples. Experiments were performed with French-Japanese spoken language texts, but the method is designed for any unexplored language pairs.

The overview of the system is immediately presented and detailed step by step in the next section. Next, the method of the experiments and the results are described, discussed, and at last, few words are given as a conclusion.

2. OVERVIEW OF THE METHOD

The flow of the method is presented in figure 1. The system accepts a new translation pair of sentences, searches multiple similar translation pairs among a set of translation examples or corpus, and computes word correspondences of the new sentence pairs, by using these selected similar pairs. If the result is not accurate, manual correction is performed. The new pair is finally appended to the corpus. On the other hand, the system learns success and failure from the results in order not to reiterate the same mistake. Alignment between two huge parallel corpus can be done by repeating the same operation for all pairs of sentences in them.

The structure of an entry in the translation examples is presented in table 1[1]. A token is represented with the format "token/POS (Part Of Speech) tag". For the tagging operation, INALF[2] (Institut

[1] Meaning of the sentence: "Do you have an ashtray?"

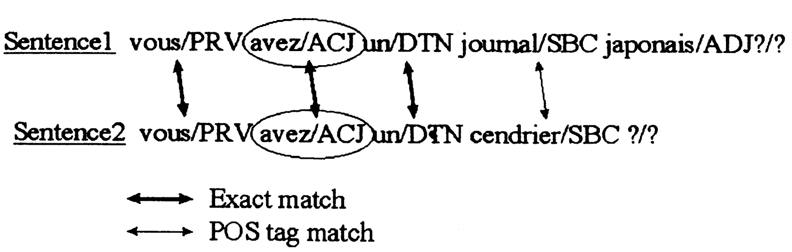
Sentence1 vous/PRV avez/ACJ un/DTN journal/SBC japonais/ADJ ?/?
Sentence2 vous/PRV avez/ACJ un/DTN cendrier/SBC ?/?
Exact match
POS tag match

Figure 2: Partial matching method

National de la Langue Française) s EBTI program was used for French sentences and CHASEN1.51[7] tagging program for Japanese sentences. INALF has proposed 70 POS tags for the French language. We removed differences between plural and singular, since they make negative contribution to capture similarity of two sentences. 48 POS tags are therefore used. 20 of them are punctuations and 3 are particular cases. As for the Japanese language, POS tags are hierarchically classified, and only the highest level consisting of 14 POS tags is considered. Utilization of syntactic analyzers or semantic analyzers might produce a better prediction of word correspondences. However, these tools themselves, are not extremely accurate. On the other hand, recently developed lexical analyzers are nearly perfect and quite available for different languages. Therefore, only POS tags were used during the prediction.

Correspondence maps are of the form $wpf1, wpf2, .../wpj1, wpj2, ...$, where $wpfi$ are word positions in the French sentence, and $wpjj$ word positions in the Japanese sentence. In the example of table 1, "2/3" means that the word "**avez**" corresponds to "*ari*". In the same way, "**cendrier**" corresponds to "*haizara*" (ashtray), and "?" to "*.*". During the manual correction word correspondences were determined according to the meaning of both sentences. Consequently, different possibilities were allowed. A token can have no correspondent, as the case of zero pronoun or some Japanese particles like "*wa*", or some French prepositions and articles. It can also have many correspondents, as in the case of "**s' il vous plait**" - "*kudasai*" (please). Besides, many tokens may correspond to many tokens, as is the case with "**voulez - vous**" - "*itadake masu ka*" (could you). In multiple tokens-involved correspondences, non-contiguous segments are also allowed, as in the case of "**n' pas**" - "*masen*" (not) of the segment "**n' ai pas**" and "*ari masen*" (do not exist). Restrictions were avoided because they make the method very specific and unable to face different situations. Inversely, a system which can support different kinds of word correspondences is able to run with any restricted and standardized situation.

3. MATCHING PROCESS

3.1 Similarity between two segments

During the matching process, multiple translation sentence pairs which match the new sentence pairs best are selected from the corpus. If two sentences are similar in one language, their translation sentences are not necessarily similar in the other language. Therefore, similarity between segments of sentences is preferred. It is more probable to discover similar pair of translation sentences if only shorter segments are observed. The search of similar segments of sentences or partial-matching processes is performed independently for both languages. For each word of an input sentence, n sentences having a segment matching the segment containing that word are selected. The partial matching algorithm is as follows:

1. For each token of the input sentence, search an exact matching token in the example sentence
2. If found, start a backward and forward comparison from the found exact match (the comparison starts with considering exact match, and continues with POS tag match when exact match expires, and stops when a mismatch is encountered.)

An illustration is presented in figure 2[2]. For the token "**avez/ACJ**" of the first sentence, an exact match is at the second position in both sentences. A backward comparison produces one exact match "**vous/PRV-vous/PRV**", and a forward one yields one exact match "**un/DTN-un/DTN**", which

[2] The meanings of the first sentence and of the second sentence are respectively "do you have a japanese newspaper?" and "do you have an ashtray?"

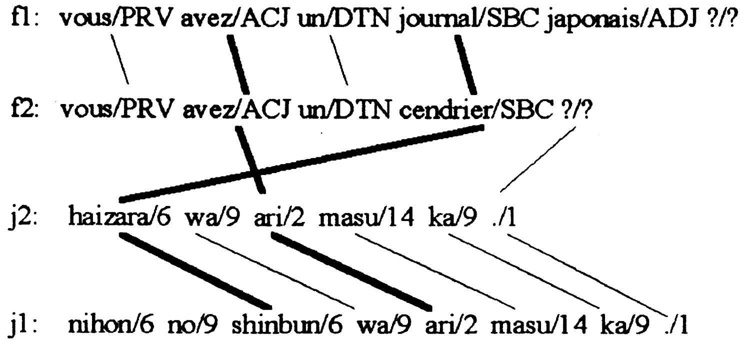
f1: vous/PRV avez/ACJ un/DTN journal/SBC japonais/ADJ ?/?
f2: vous/PRV avez/ACJ un/DTN cendrier/SBC ?/?
j2: haizara/6 wa/9 ari/2 masu/14 ka/9 ./1
j1: nihon/6 no/9 shinbun/6 wa/9 ari/2 masu/14 ka/9 ./1

Figure 3: Search of the word correspondence by analogy

is followed by one POS tag match **"journal/SBC-cendrier/SBC"**. To select the n best matching segments for each token, the following similarity metric is used:

$$SM = \alpha * NE + NP \tag{1}$$

where SM is the similarity metric, NE the number of exact matches and NP the number of POS tag matches. For the sentences of figure 2, there are 3 exact matches and 1 POS tag match. The similarity metric is therefore:

$$SM = \alpha * 3 + 1 \tag{2}$$

3.2 Similarity between two pairs of segments

To predict the word correspondences between a new pair of sentences, similar pairs are required. The selected similar sentence will be rejected if any similarity cannot be found between its translation sentence and the translation of the source sentence. The above-mentioned partial matching method is performed independently for both languages. Similarity between translation of the selected sentence and the other language-part of the new pair of sentences has not yet been confirmed. The same algorithm is applied again here to verify that similarity.

It is important to mention that the number of segments selected, which were declared to be n per token, may change depending on the situation. No segment is selected when any exact match for the token does not exist. Besides, a same similar segment might be selected for different tokens. In other words, n is the maximum. At most $6n$ similar segments are selected for 6 tokens sentence.

4. PREDICTION OF THE WORD CORRESPONDENCES OF THE NEW TRANSLATION EXAMPLE

After having selected the similar pairs of segments, word correspondences between the new pair of sentence is predicted. The process is described in figure 3. $(f1, j1)$ is the new pair of sentences. $(f2, j2)$ is a selected pair from the translation examples. Correspondences between $f2$ and $j2$ are defined beforehand within the translation examples. Correspondences between $f1$ (resp. $f2$) and $j1$ (resp. $j2$) are determined during the matching process. The main task here is to find all paths starting from an element of $f1$ and reaching an element of $j1$. In this case, there are two paths, which correspond to two word correspondences **"avez/ACJ**-*ari/2*" and **"journal/SBC**-*shinbun/6*".

Similarity metrics which where computed during the comparison of source sentences and of translation sentences are added. The prediction process starts with the similar pair having the highest value of that sum. Correspondences involving multiple tokens, which are many-to-many or one-to-many links, are considered only when they came from the same similar pair of segments. In other words, if a token has already been involved in a previous correspondence, any following correspondence involving it will be rejected unless the same similar segment was used during their prediction.

Table 2: Data for the experiments

Total number of translation examples	2,600 pairs
Average length of Japanese sentence	7.74 tokens
Average length of french sentence	7.84 tokens
Average number of word correspondences per sentence	7.27

5. FEEDBACK LEARNING PROCESS

The system is designed to be operated by different users. Sentences from different domains and incorrect sentences may be typed in. Different points of view, as well as mistakes may also appear during the manual correction. The feedback learning process is introduced here to support the above-mentioned situations. Its idea has already been proposed [3]. Each example in the corpus is given a numeric value FP (Feedback Parameter). If the prediction is not correct, the FP of the involved similar example will be decremented by 1. If it is correct, the value will be incremented by 1. The value FP is combined with the similarity metric to determine the examples which will be use to predict word correspondences of subsequent pairs. Two conditions are defined on FP:

1. Decrementation is always applied on FP when the prediction results are not correct. However, incrementation is applied only to FP whose value has not returned to the initial value. This is decided not only to eliminate incorrect translation examples, but also to avoid a possible restriction to particular translation examples having a high value of FP if incrementation is always considered.

2. FP is initialized to -1. The higher its absolute value is, the more the rejection of the example is probable. During the search of the similar pair the following metric is used:

$$NM = SM/abs(FP) \tag{3}$$

SM is the similarity metric, and NM the new metric. NM is not a similarity metric but a metric which is used to decide which sentence is better to predict word correspondences of subsequent pair of sentences.

6. EXPERIMENTS AND RESULTS

The corpus which was used during the experiments is described in table 2. Sentences were taken from French-Japanese conversation books [8, 9]. The presence of short sentences in the conversation language world reduces the average length of a sentence. The translation examples base is initially empty. Sentences are entered pair by pair into the system which predicts the word correspondences between them. Results are corrected manually if necessary, and finally the new pair is inserted in the corpus.

α of equation (1) is set to 10, and the equation of the similarity metric becomes:

$$SM = 10 * NE + NP \tag{4}$$

If α is set to high value the similarity metric almost depends on the number of exact matches or NE. The effect of the number of POS tag matches appears only when the numbers of exact matches are almost the same. n is set to 5. In other words, at most 5 segments are selected for each token of the input sentence.

Preliminary experiments were carried out to determine these values of α and n. 800 pairs of translation examples were used and 200 new pairs were entered into the system. Different values ranging from 1 to 10 were given to α and to n. The exact prediction rate and the ratio of the number of extracted word correspondences are observed for each situation. According to that result, the greater α is, the more high prediction rate is obtained. As for n, low value n gives higher rates. On the other hand, the variation of α does not affect the ratio of the exctracted word correspondences much. High value n gives higher ratio. The difference seems important between $n = 1$ and $n = 2$, but gets smaller as n increases. $\alpha = 10$ and $n = 5$ were the most appropriate for high prediction rates with many extracted word correspondences.

Having these parameters determined, 2,600 translation pairs were entered one by one into the system. If the predicted word correspondences are not correct, manual correction is performed. To decide whether

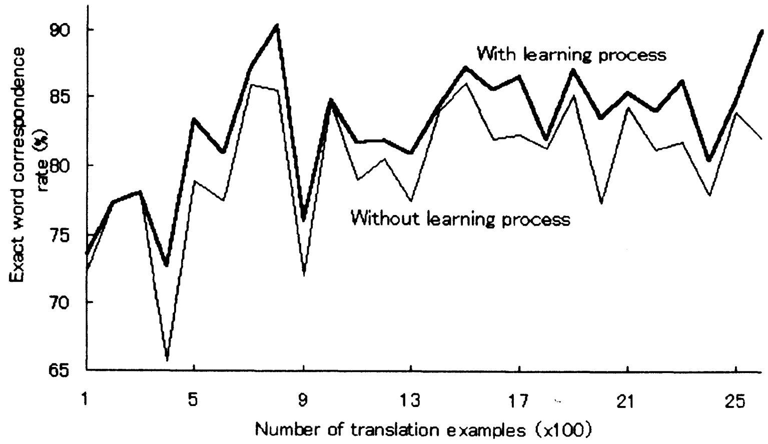

Figure 4: Variation of the exact word correspondence rate by the number of translation examples

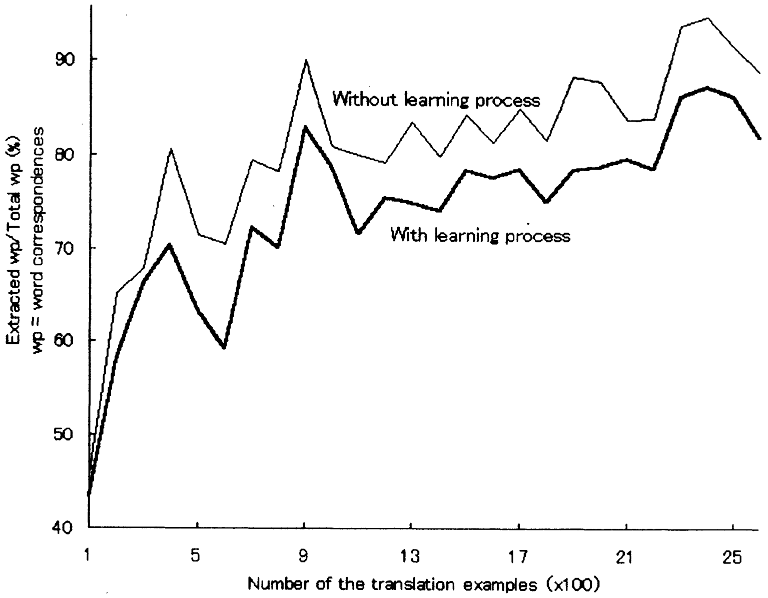

Figure 5: Percentage of the number of extracted word correspondences

the result is accurate or not we have created a list of correct word correspondences for each pair. Results are compared to the list and any missing link is considered to be false. Results are divided per 100 pairs. The correct correspondence rate, and the ratio of the number of extracted links were observed. In addition, we had done another experiment in which the learning process was not performed, and from which the effect of the learning process was verified. The exact correspondence rates and the ratio of the extracted correspondences are respectively presented in figure 4 and figure 5.

Next, the number of correspondences which were extracted from exact matches and those which were from POS tag matches were counted. It is summarized in table 3.

7. DISCUSSION

In figure 4, the accuracy rate increases as the number of translation examples increases. The slope is small but the graph does not decrease. As the number of translation examples goes beyond 1,000, more than 80% of accuracy rates are obtained. The highest value, which is 90.0%, is detected at 2,600 translation examples. The graph surpasses the one of "without learning process". It confirms the effectiveness of the learning process in the method. In figure 5, the increase of the number of the extracted links is

Table 3: Usefulness of the POS tags

	Extracted word correspondences	Exact word correspondences	Exact word correspondences rate
POS Tag Match	33.1%	26.0%	65.2%
Exact Match	66.9%	74.0%	91.9%
Total or Average	100.0%	100.0%	83.1%

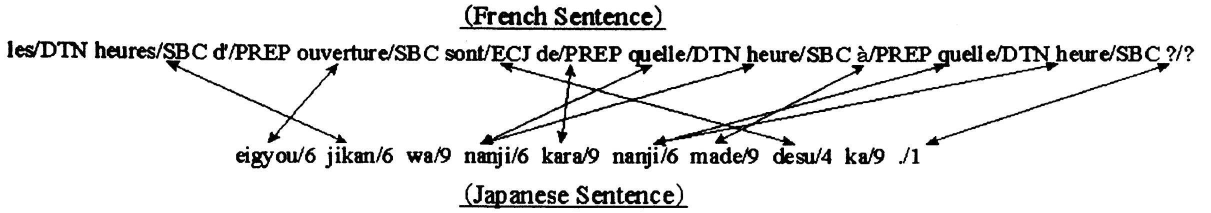
(French Sentence)
les/DTN heures/SBC d'/PREP ouverture/SBC sont/ECJ de/PREP quelle/DTN heure/SBC à/PREP quelle/DTN heure/SBC ?/?
eigyou/6 jikan/6 wa/9 nanji/6 kara/9 nanji/6 made/9 desu/4 ka/9 ./1
(Japanese Sentence)

Figure 6: Example of result

clearly visible. The graph is lower than the one of "without learning process". However, since translation patterns can still be extracted with fewer word correspondences, the accuracy rate is more important than the number of extracted word correspondences. The highest value, which is 87.2%, corresponds to 2,400 translation examples. Despite of the small number of translation examples and the presence of sentences not following grammar rules in the spoken language, good results were earned. The rising trend of both graphs, with high values being manifested, confirms the effectiveness of our method.

An example of the result is presented in figure 6[3]. As the link between **"quelle heure"** (what time) and ***"nanji"*** shows, links involving multiple tokens were successfully predicted. Logically, **"quelle"** (what) corresponds to *"nan"*, and **"heure"** (time) to *"ji"*. However, since *"nanji"* is a single token, links between **"quelle"** and *"nanji"*, and again between **"heure"** and *"ji"*, were recognized. Predictions of such links becomes successful because similar cases exist in the translation examples. In addition, links between tokens appearing more than once in a same sentence were predicted successfully. As the example of figure 6 shows, **"quelle heure"** and *"nanji"* appear twice in each sentence, but their respective correspondents are estimated accurately. Such cases can be predicted because not a single token at a time, but longer segments were observed. This cannot be carried out with statistical methods which consider single tokens as unit of observation.

According to table 3, 33.1% of the extracted word correspondences are from POS tag matches. This means that in spite of the absence of an exact match in the selected similar segments, correspondences were predicted successfully by POS tags. 65.2% of accuracy rates are still far from perfect, but it is very promising. The ratio of the number of extracted links would have dropped from 75–80% tó 50–55% if POS tags were not used. That confirms the usefulness of POS tags.

The number of POS tags of both languages affects the results of the matching process. If few POS tags are used, long similar segments can be discovered. However, if that number is too small, incorrect links easily arise, especially with high frequency POS tags. On the other hand, the greater the number of POS tag is, the less a long similar segment can be found. In that case, links between unregistered tokens are hard to discover. That was one reason of failure. In contrast to Japanese language' s 14 POS tag, French language' s 40 POS tags needs some adjustments.

During the matching process, tokens, which could not be aligned with any tokens of similar segments, cause failure. If any link cannot be discovered within the selected similar segment, the prediction process cannot be performed. On the other hand, for links involving multiple tokens, the links are accepted if the prediction is done by a single similar segment. However, if other segments must be used, only the links which can be obtained by the first single segment are approved, and the links becomes incomplete.

[3] In english, the sentence means "From what time to what time are the business hours?"

Enlargement of the translation examples will of course solve these problems. Nevertheless, as an improvement of the method, study on covering the whole input sentence with the selected examples, as well as on restoration of splitted links involving multiple tokens, is still a necessity.

8. CONCLUSION

In this paper, we have described a method for aligning a pair of translation sentences at sub-sentential level and by analogy. Statistical methods cannot produce reliable results when the corpus size is limited. Therefore, they cannot be applied to unexplored languages whose huge parallel corpus is not available. A method which uses a word correspondences-included initial translation examples is proposed. Links between new pairs of translation sentences is predicted according to links between similar examples. By doing so, links involving multiple tokens, as well as links between a token appearing more than once in the same sentence, can be predicted. Besides, introduction of the learning process avoids reiteration of a previously encountered failure.

The rising trend of the number of extracted links, by the augmentation of the translation examples was confirmed. The accuracy rate of prediction is also slightly rising. With 2,600 translation examples, more than 80% of the word correspondences are extracted, with a 90% of accuracy rate of prediction. This is considered to be a good result as far as spoken-language sentences are concerned. On the other hand, the effectiveness of the learning process is confirmed by the fact that the graph of the accuracy rate of prediction surpasses the one without learning process. In addition, links involving multiple tokens and respective links of token appearing more than once in a same sentence were predicted successfully.

Adjustment of the number of POS tags in both languages, covering the input sentence with the selected examples during the matching process, and restoration of splitted links involving multiple tokens is the next direction of this study.

Acknowledgement: This work is partially supported by the Grants from the Government subsidy for aiding scientific researches (No. 10680367) of the Ministry of Education of Japan. Thanks to the "Institut National de la Langue Françise" for providing the EBTI tagger program for the French language.

9. REFERENCES

[1] T. Andriamanankasina, K. Araki, Y. Miyanaga and K. Tochinai: "Method for Searching the Best-Matching Sentence in Example-Based Machine Translation", Technical Report of IEICE, Vol. NLC97-10, pp. 15–20, Japan, 1997.

[2] T. Andriamanankasina, K. Araki, Y. Miyanaga and K. Tochinai: "Machine Translation Based on the Relations between Words", Proc. of Towards Useful Natural Language Procesing Symposium, Japan, 1997.

[3] K. Araki, Y. Takahashi, Y. Momouchi, and K. Tochinai : "Non-Segmented Kana-Kanji Translation Using Inductive Learning", Transactions of the IEICE, Vol. J97-D-II, No. 4, pp. 391–402, 1996.

[4] P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer: "The Mathematics of Statistical Machine Translation: Parameter Estimation", Computational Linguistics, Vol. 19, No. 2, pp. 263–309, 1993.

[5] D. Melamed: "A Word-to-Word Model of Translational Equivalence", 35th Conference of the Association for Computational Linguistics ACL'97, Spain, 1997.

[6] M. Kitamura, and Y. Matsumoto: "Automatic Extraction of Translation Patterns in Parallel Corpora ", Transactions of the IPSJ, Vol. 38, No. 4, pp. 727–736, 1997.

[7] T. Yamashita: "ChaSen Technical Report", Nara Advanced Institute of Science and Technology, 1996.

[8] S. Meguro: "Manuel de Conversation Française", Hakusuisha, Tokyo, 1987.

[9] F. Sato: "Locutions de base", Hakusuisha, Tokyo, 1990.