Example-Based Machine Translation of Part-Of-Speech Tagged Sentences by Recursive Division

Tantely Andriamanankasina, Kenji Araki and Koji Tochinai
Graduate School of Engineering, Hokkaido University
Kita 13 Nishi 8, Kita-ku, Sapporo 060-8628, Japan
Email: {tantely,araki,tochinai}@media.eng.hokudai.ac.jp

Abstract

Example-Based Machine Translation can be applied to languages whose resources like dictionaries, reliable syntactic analyzers are hardly available because it can learn from new translation examples. However, difficulties still remain in translation of sentences which are not fully covered by the matching sentence. To solve that problem, we present in this paper a translation method which recursively divides a sentence and translates each part separately. In addition, we evaluate an analogy-based word-level alignment method which predicts word correspondences between source and translation sentences of new translation examples. The translation method was implemented in a French-Japanese machine translation system and spoken language text were used as examples. Promising translation results were earned and the effectiveness of the alignment method in the translation was confirmed.

1 Introduction

In the traditional Rule-Based Machine Translation (RBMT), huge amount of rules and dictionaries need to be prepared and maintained [1]. To avoid that hard and time-consuming task, Example-Based Machine Translation (EBMT) was proposed [2], The basic idea of EBMT is to extract, among a collection of translation examples, a number of translation examples whose source sentence is similar to the sentence to be translated and achieve the translation task by imitating these translation examples. Rules are not required. Instead, the system learns from translation examples. Various EBMT models have been proposed [3. 4. 5] and different issues were discussed [6].

Current EBMT models use syntactic analyzer results. Utilization of syntactic and semantic analyzers is expected to produce accurate translation results. However, these tools themselves are not perfect and still not available in languages where research has not been sufficiently carried out. On the other hand, recent lexical analyzers [7, 8] are extremely precise. We therefore, have proposed an EBMT method not depending on syntactic or semantic analyzers [9]. Lexical analyzer is merely used together with the parallel and aligned corpus. The study is limited to Part Of Speech (POS) tags. However, the translation model is considered to be adaptable to any additional tags which may raise the precision of the translation. If tagged text are available, or the recently proposed Global Document Annotation[1] becomes widespread, highly accurate translation system will be expected. Our proposed method is implemented in a French-Japanese EBMT system. However, it is generally designed for languages whose resources are hardly available.

The translation is possible and a correct result can be earned when the input sentence is almost covered by the matching sentence. However, when such example cannot be discovered, it is hard to translate sentences correctly. Even for relatively short sentences, there are cases where uncovered segments, when they exist, cause errors. Besides, there are errors from mismatches appearing during the matching process. To solve these problems, we propose in this paper a translation method which is based on a recursive division of the input sentence and translate each part independently. The system predicts the position where the input sentence or a segment of the input sentence should be divided according to the links existing between the source sentence and the target sentence of the extracted translation example.

On the other hand, link-included translation examples are used. Links are word-to-word correspondences between a source sentence and a target sentence. They are especially necessary to extract translation patterns from translation examples. For example, consider the French input sentence “**je suis heureux**[2] (I am happy)”', and suppose that the sentence “**je suis malade** (I am ill)” is the best match

[1] http//www.etl.go.jp/etl/nl/GDA/

[2] French words are presented with **boldface** characters

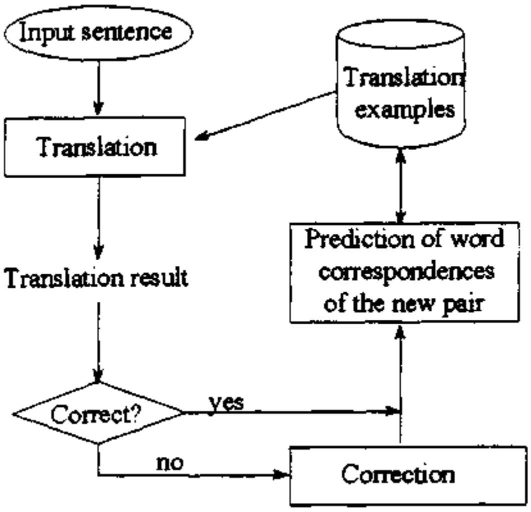

Figure 1: Overview of the translation system

Its Japanese translation is *"watasi ha byouki desu"* [3]. If the link between **"malade"** and *"byouki"* can be identified, the translation result, which is *"watasi ha siawase desu",* will easily be obtained by replacing *"byouki"* by the translation of "**heureux",** which is *"shiawase",* in the translation of the best match.

Current methods for such alignment of parallel text at word level are all based on statistics [10, 11, 12]. Statistical methods are not able to produce reliable result with size-limited corpus. Besides, prediction of links involving multiple tokens or links of a token appearing more than once in a sentence, has not entirely been resolved. To solve these problems, we have proposed a analogy-based word level alignment method using a link-included initial translation examples [13]. Experiments confirmed that with 2,400 translation examples, more than 80% of the links are extracted with 90% of accuracy rate of prediction. Besides, the accuracy rate is rising as the number of translation examples multiply. This alignment method was introduced in the translation system to predict links of new examples. This makes the translation system able to use new examples. An evaluation of the application of this alignment method in the translation system is also presented in this paper, in addition to the description of the translation method.

The translation system is immediately presented and detailed step by step in the following sections. The second part of the paper describes the experiments, results and discussions.

2 Overview of the translation system

The overview of the translation system is presented in Figure 1. The translation itself is performed using a number of translation examples whose source sentences match the input sentence. If the result is not correct, the system asks the user to input the correct translation result. Links between the input sentence and the correct translation result will be predicted, again using similar translation examples, which are extracted from the parallel corpus. The new translation example, link included, is finally appended to the corpus.

An entry in the translation examples is presented in Table 1. It is composed by the French source sentence, its Japanese translation sentence, and a map describing links between words in both sentences. A token is presented with the format "token/POS tag". For the tagging operation, INALF's[4] EBTI program was used for French sentences and CHASEN1.51 [7] tagging program for Japanese sentences. EBTI is an adaptation of the Eric Brill Tagger [8] for French. There are 48 POS tags for French language, and 14 for Japanese language.

A link has the form "Wf_1, Wf_2, ../Wj_1, Wj_2,..". "Wf_i" are word positions in the French sentence and "Wj_i" word positions in the Japanese sentence. In the example of Table 1, "2/2" means that the token **"suis** (be)" corresponds to "*desu*". By the same way. "3,4/1" means that "**sans profession** (jobless)" corresponds to *"musyoku".* Words or segments of words having no correspondent are not specified. For the reason being described in section 3, only links which are formed by contiguous words are considered. For example, to align the phrase **"ne va pas** (do not go)" with "*iki masen",* the obvious way might be aligning **2ne pas** (do not)" with *"masen"* and "**va** (go)" with *"iki".* However, since "**ne pas**" is not a contiguous segment, only one link between "**ne va pas"** and *"ikimasen"* is considered. Although "**va**" alone cannot be translated, a contiguous segment map is obtained. This requirement should not raise problem since non-contiguous segment map can always be modified to a contiguous one by combining segments.

3 Translation method

A simple illustration of the translation idea, is presented in Figure 2. The input sentence and matching sentence mean "Jean is seriously ill" and "He is rich" respectively. A translation example having a source sentence matching the input sentence is extracted from the translation example. The input sentence is divided at the position of the word "**est** (be)", which is a common segment for both sentences. In the source sentence of the selected example, the segment "**il** (he)" is located on the left side of the common segment and **"riche** (rich)" on his right side. If one observes their correspondents in the target sentence, the structure "(left side) *ha* (right side) *desu"*

[3] Japanese words are presented with italic characters

[4] Institut National de la Langue Française

Table 1: Structure of a translation example

French Sentence	je/PRV suis/ECJ sans/PREP profession/SBC
Japanese Sentence	*musyoku*/6 *desu*/4
Links	2/2 3,4/1 5/6

PRV: pronoun, PREP: preposition, SBC: common noun, ECJ: verb "etre"
6: noun, 4: assertive

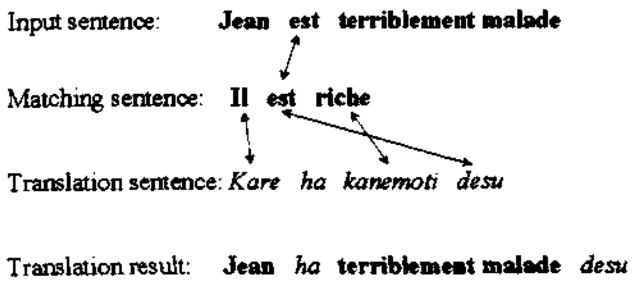

Figure 2: Illustration of the idea of the translation method

of the target sentence will be discovered. This structure will be applied to the input sentence, and finally it can be rewritten as **"Jean** *ha* **terriblement malade** *desu".* Using other examples, the same process will be applied again and again to non-translated segments, while they exist. In other words, the recursion is the application of the same process to the result of its previous execution. Here, there are two non-translated segments **"Jean"** and **"terriblement malade". "Jean"** is a one word segment and can be translated without division. On the other hand, **"terriblement malade"** can again be divided into **"terriblement"** and **"malade"** by the same process, or be translated directly if it appears somewhere in the corpus. Whether they need a division or not, appropriate examples must be selected and applied to translate them.

The present method has 2 important advantages. First, utilization of syntactic analyzer is unnecessary because sentences can be translated without understanding their syntactic structures. Second, any sentence can always be divided. This characteristics makes the method able to translate long sentences. If non-translated segments can be divided at the right position at every step, the correct translation result will be reached.

There are therefore two main steps in the translation method.

1. The extraction of examples having a source sentence similar to the input sentence, and
2. The production of the translation result.

They are described in details in the following sections.

3.1 Extraction of examples

According to the idea of the method, sentences having a common segment with the input sentence will be the target of the extraction. On that segment, the sentence or segment of sentence will be divided into independent parts. Similarity in the structure of both sentences must also be considered because it assures the similarity between both right side segments and between both left side segments.

A condition is imposed to the common segment. In the translation example, it must be linked with the target sentence by one-to-one contiguous segment link. Consider the source sentence **"nous sommes amis.** (We are friends.)" and its target sentence *"watasitati ha tomodati desu".* **"nous** (we)", **"sommes** (be)", "**amis** (friends)", and "." correspond to *"watasitati", "desu", "tomodati"* and "." respectively. If **"nous sommes"** exists in the sentence to be translated, it will be a common segment. However, since the correspondents of "**nous",** which is *"watasitati",* and **"sommes"**, which is *"desu",* are separated in the target sentence, the translation will be performed considering only either "**nous"** or **"sommes"** as the common segment. Otherwise, the sentence could not be divided.

On the other hand, in the case where a segment having no link is the common segment, the whole segment is considered as fixed segment and must match exactly with their correspondents. For example, consider the source sentence "**je vous remercie.** (I thank you)" and its translation *"arigatou gozaimasu*". Here, "**remercie** (thank)" corresponds to *"arigatou"* and "**je** (I)" as well as **"vous** (you)" have no correspondent. Therefore, since "**je"** and "**vous"** are contiguous, not only one but both of them must exactly match their correspondent in the input sentence, when it is the common segment. Otherwise, this translation example will not be selected.

To cover the input sentence, a number of examples are extracted. The matching algorithm is as follows.

1. For each token of the input sentence, search a same token in the source sentence.
2. If found, from that position, start a forward and backward search of matches. The search starts with exact matches and continues with POS tag matches when a POS tag match or a non-contiguous match is encountered.

To select the best matching sentences, the following similarity score is used.

$$SC = \alpha \times NE + \sum_{i=1}^{NP} \frac{2}{D_i} \qquad (1)$$

SC is the similarity score. *NE* the number of exact matches. *NP* the number of POS tag matches and D_i the distance separating a token from the common segment, measured in number of tokens. This equation of similarity score has been proposed and discussed in details in [13], with a description of the preliminary experiments for fixing the value of *a*. According to it, the suitable value of *a* is 10. This value makes heavier the presence of exact matches, which are located within or around the common segment, compared to the POS tag matches. In [13], since non-contiguous matches were not considered, D_i was ignored and the number of POS tag matches was only taken into account. However, in the present method, search for matches continues until it reaches the head or the end of the sentence. Therefore, D_i is introduced to make the difference between matches located near the common segment from those located far from it. The value 2 of the dividend is explained by the presence of 2 sentences being involved in the comparison. One sentence is selected for each token of the input sentence. It is the sentence having that token as part of the common segment and having the highest value of similarity score.

It is important to note that D_i never becomes 0 because it concerns only POS tag matches. POS tag matches are located outside the exact matching segment. There must therefore be a distance, at least being equal to 2 because 2 sentences are involved in the comparison, separating any POS tag match and the exact matching segment.

This algorithm, since it always starts the search from an exact match, has a subsidiary advantage that processing time can be reduced considerably by indexing the corpus on each token.

An illustration is presented in Figure 3. Sentence 1 and 2 mean "Do you have a Japanese newspaper?" and "Do you have an ashtray?" respectively. For the token **"avez/ACJ"** of sentence 1, an exact match is detected at the second position in both sentences. A backward search produces one exact match, "**vous/PRV-vous/PRV**". A forward search yields one exact match, "**un/DTN-un/DTN".** The following, **"journal/SBC-cendrier/SBC"** is not an exact match. Therefore, from this position, only POS tags are observed. That makes the match between "?/?" and "?/?". although it is an exact match, to be considered as POS tag match. This consideration strengthen matches between the common segments which are located within the exact match segments. There are 3 exact matches and 2 POS tag matches. The common segment is "**vous/PRV avez/ACJ un/DTN**". As far as the sum of distances separating token from the common segment in both sentences is considered, the first POS tag match are away by 2 tokens, one in each sentence, and the second by 5 tokens. 3 in the first sentence and 2 in the second one. Consequently, the similarity score is as follows.

$$SC=10x3+\frac{2}{2}+\frac{2}{5}= 31.4$$

Here, the token "**japonais/SBC"** is skipped because any correspondent does not exist. In the case where multiple interfering matches are discovered, the match which is close to the last match is selected and the search continues.

3.2 Production of the translation result

During the search of similar sentences, one translation example is extracted for each token of the input sentence. However, since a same sentence may be extracted for multiple consecutive tokens and exact match cannot be discovered for unregistered words, there are cases where the number of extracted examples is fewer than the number of tokens. The translation result is produced using these examples. Consider the input sentence "**nous sommes camarades d' école** (We are schoolmates)". The flow of the production process is illustrated in Figure 4. Example 1 and example 2 mean "We are friends from long ago" and "He is my childhood friend" respectively. In the first translation example, **"sommes** (be)" is considered as the common segment. By observing the correspondents of its left side segment and of its right side segment in the target sentence, the pattern "(left-side) *ha* (right-side) *desu"* can be extracted. An application of this pattern to the input sentence produces **"nous** *ha* **camarades d' école** *desu".* Here *"ha"* has no correspondent. However, since it is an element which is located in the middle of the right side and left side segments, it remains. The absence or presence of these tokens having no correspondent, like *"ha"* in the present case, sometimes modify completely the translation result. Two cases where segments having no correspondent are kept are proposed.

1. They are located between the translation of the right side and one of the left side segments. Since the common segment is located between the left-side and right-side segments, it is assumed that a segment which is located between the translation of the right side and one of the left side segments in the target sentence plays an important role when the common segment exist, as the case of "ha" in the above example.

2. They are closely related to the translation segment of the common segment (prefixes, postfixes, particles). This is the most obvious case since if they depend on the translation of the common segment, they should automatically follow it.

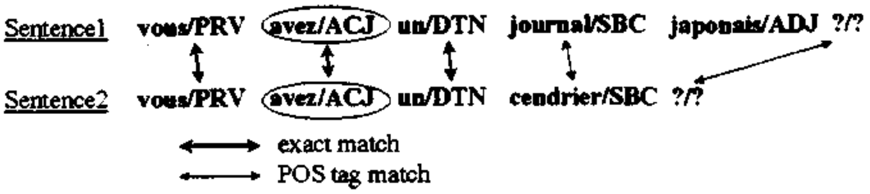
Sentence1 vous/PRV avez/ACJ un/DTN journal/SBC japonais/ADJ ?/?
Sentence2 vous/PRV avez/ACJ un/DTN cendrier/SBC ?/?
exact match
POS tag match

Figure 3: Illustration of the matching method

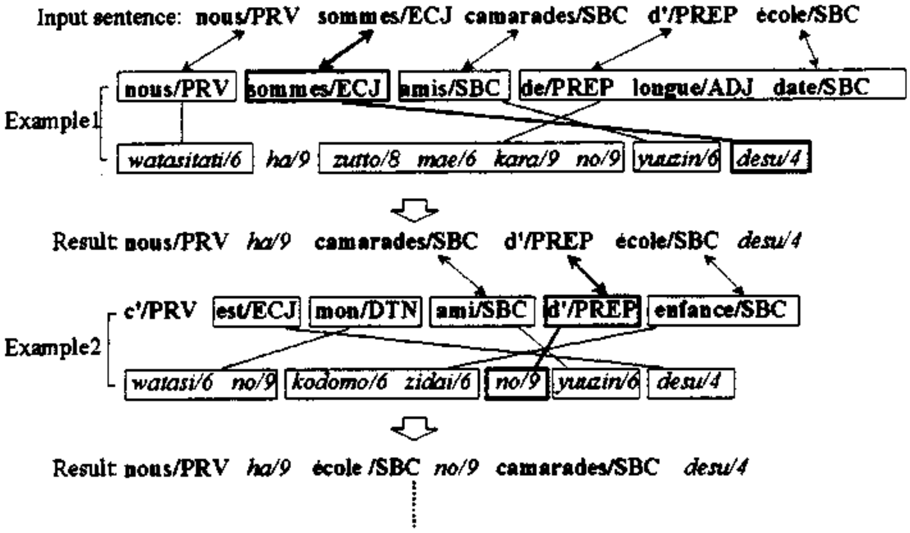
Input sentence: nous/PRV sommes/ECJ camarades/SBC d'/PREP école/SBC
Example1
nous/PRV sommes/ECJ amis/SBC de/PREP longue/ADJ date/SBC
watasitati/6 ha/9 zutto/8 mae/6 kara/9 no/9 yuuzin/6 desu/4
Result: nous/PRV ha/9 camarades/SBC d'/PREP école/SBC desu/4
Example2
c'/PRV est/ECJ mon/DTN ami/SBC d'/PREP enfance/SBC
watasi/6 no/9 kodomo/6 zidai/6 no/9 yuuzin/6 desu/4
Result: nous/PRV ha/9 école /SBC no/9 camarades/SBC desu/4

Figure 4: Illustration of the production of the translation result

After the application of example 1, the output contains 2 non-translated segments "**nous**" and **"camarades d' école".** The recursion continues with the application of example 2 to the segment **"camarades d' école"** by the same method. Only the matching segment **"ami d' enfance"** is therefore considered. "**d'**" is the common segment. According to the position of the translations of their left side segment **"ami"** and right side "**enfance", "camarades d' école"** can be rewritten as "**école *no* camarades**". After the application of example 2, there still are 3 non-translated segments **"nous", "école" and "camarades"**. The translation process continues with the translation of these segments by the same method. Here, the case is specific because segments are each composed by 1 token. They can be translated directly without segment division.

In addition, there are cases where left side or right side segment disperses. Consider the translation example 2. The segment "**d'** (of)" is considered as the common segment. The correspondents of its left side segment in the target sentence are *"watasi no* (my)", *"yuuzin* (friend)" and *"desu* (be)". And its right segment corresponds to *"kodomo zidai* (school days)". Therefore, a pattern like "(left side) (right side) *no* (left side)" will be extracted from the target sentence. Two "(left side)" segments appear in this pattern and the sentence cannot be divided. In that case, priority is given to the one which is close to the common segment. Here, *"yuuzin"* is selected and the extracted pattern will be "(right side) *no* (left side)".

Actually, if translation example 2 is used before translation example 1, that means if it is applied to the initial input sentence, it will produce an incorrect result. It produces a result like "**école** *no* **nous sommes camarades**". This surely generates a completely different final result like *"gakkou no watasitati ha yuuzin desu"*. In short, the order of use of the extracted translation examples is very important. We propose three conditions to decide this priority order.

1. Top priority is given to examples having common segments which divide successfully the sentence without dispersion of each part. Punctuation, conjunctions and so forth generally fall into this category.

2. Next, examples having common segments containing a verb are considered. This is explained by the importance of verbs in recognizing the structure of the whole sentence. It is assumed that the corpus is large enough to contain different sentences including a given verb. Otherwise, the matching method would not produce examples having the same structure as the input

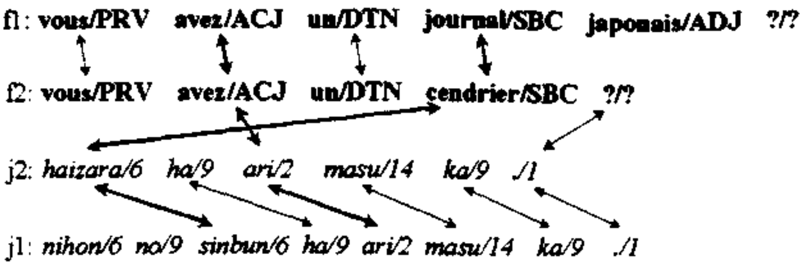
f1: vous/PRV avez/ACJ un/DTN journal/SBC japonais/ADJ ?/?
f2: vous/PRV avez/ACJ un/DTN cendrier/SBC ?/?
j2: haizara/6 ha/9 ari/2 masu/14 ka/9 ./1
j1: nihon/6 no/9 sinbun/6 ha/9 ari/2 masu/14 ka/9 ./1

Figure 5: Prediction of links of a new translation example

sentence.

3. For the rest, the similarity score will make the difference on condition that examples having non-functional words like nouns, adverbs or adjectives as common segment will be the last to be considered.

4 Prediction of links of the new pair

There are two steps in the prediction of links between the input sentence and the correct translation result.

1. The search of examples to be used for the prediction, and
2. The prediction process itself

For the first step, examples whose source sentences are similar with the input sentence and target sentences with the correct translation result, will be extracted. 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 pairs of translation sentences if only short segments are observed. The search algorithm is similar to the one described in section 3.1 except that skip is not allowed and the search stops when a mismatch is encountered. This means that in the case of Figure 3, since **"japonais/ADJ"** is skipped, link between "?/?" and "?/?" is rejected. In addition, the distance D_i is ignored and the similarity score is as follows.

$$SC = 10 \times NE + NP \quad (2)$$

For the example of Figure 3. since one POS tag match is rejected, there are 3 exact matches, and 1 POS tag match. The similarity score is therefore:

$$SC = 10 \times 3+1 = 31$$

The search of similar sentences is performed separately for both languages. For each token, at most five sentences having an exact match on that token are selected. Translation examples whose source sentence and target sentence do not simultaneously appear in the selected sentences, are rejected. The rest will be the similar translation examples. The sum of similarity scores of each component (source and target sentence) is taken as the similarity score of the translation example.

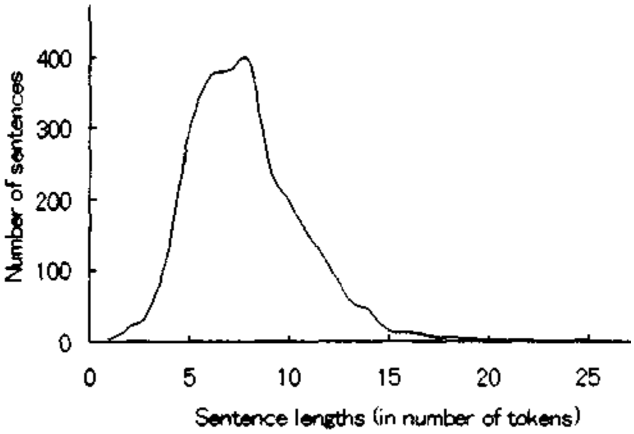

Figure 6: Distribution of French sentence lengths

As for the link prediction itself, the process starts with the example having the highest similarity score. An illustration of the method is given in Figure 5. f1 and j1 are the new translation examples whose links are to be predicted. f2 and j2 are the link-included similar examples, which were extracted from the translation examples. 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 links "**avez/ACJ**-*ari/2*" and "**journal/SBC**-*sinbun/6*". Links of segments which were not covered will be predicted with the following extracted similar examples. We encourage the reader to consult [13] for further details of the link prediction method .

5 Experiments and results

The initial bilingual corpus was composed by 2,500 examples. Sentences were taken from French-Japanese conversation books [14, 15]. The average length of sentences are 7.74 tokens for Japanese and 7.84 for French. To give the reader a better understanding of the data being used, the distribution of French sentence lengths is presented in Figure 6. New 400 French sentences, taken from the same sources, are entered one by one into the system to be translated.

First, as soon as the translation result comes out, the new translation example is not appended in the corpus. The initial 2,500 examples are used to translate all these 400 input sentences. We call it "Experiment 1". It was carried out to be able to compare the result with the case where the corpus is incremented. Next, the translation result is corrected if necessary, links between words of the new pair are predicted, and the new example is appended to the corpus. We call it "Experiment 2". To avoid a possible degradation of the whole system, newly appended examples are prevented from being involved in the prediction of links of new translation examples. They are considered only during the translation process itself.

Since unregistered words exist and a dictionary is not used, French words sometimes remain within the translation result. Evaluation of such results by sight is very difficult. We focused on segment position and consider the translation as correct if it has the same structure as the correct translation and all segments are put at their right position. The 400 input sentences are divided per 50 sentences and the result is presented in Table 2. "Ratio of use of new examples" in column 6 represents the ratio of the new translation examples in the extracted translation examples. In addition. 10 sample results are selected randomly from the output and presented in Table 3.

6 Discussions

In Table 3, the correct translation rate is 62.0% in "Experiment 1". It rises to 68.5% in "Experiment 2". Despite of the small number of translation examples and the presence of sentences not following grammar rules in the spoken language, these are considered to be very promising results. The non-rising trend of the values in "Experiment 2", where the corpus is incremented, can be explained partially by the difficulties in the translation of long sentences compared to short sentences. As being noticed in the variation of the correct translation rate and the variation of the average length of input sentences, more short sentences are correctly translated than long sentences.

However, the difference between the correct translation rates in "Experiment 1" and in "Experiment 2" in column 5. shows the effect of the utilization of new translation examples. That difference slightly drops at the 101th and at the 301th input sentences, but it generally increases as the number of translation examples increases. This confirms the effectiveness of the link prediction method in the translation system. It is necessary to verify if these new translation examples were really involved during the production of these translation results. That is the purpose of the column 6. At first, since the corpus does not contain any new translation examples, but is composed solely by the initial corpus, low values of that ratio are manifested. However, the rising trend of that ratio is clearly visible as the number of translation examples increases.

Table 3 shows examples where the input sentence is successfully divided. For example, the third input sentence "**il m' a écrit qu' il avait neigé la veille**" is divided at the position of "**qu'**". It produces the pattern "**il avait neigé la veille** *to.* **il m' a écrit**". which leads to the result *"zenjitu* **(avait neigé)** *to. kare ha kaite kita".* On the other hand there are cases of failure, as in the input sentence 4. The structure of the result is completely different from the correct translation. This failure is a result of a wrong order of the examples which were applied during the generation of translation result. Of course it needs an improvement, but we emphasize the worthiness of the correct translation rate with the defined priority order.

Besides, there are cases where the translation of some words, although they are necessary, do not appear in the result. For example in the input sentence "**demain soir, madame S donnera un bal chez elle**", "**madame**" and "**chez elle**" disappear but their translations are not present. These words have no correspondent in the target sentence of the selected examples, or their correspondents have not been predicted. With the present link prediction method, reliable translation results are earned. However, improvements still have to be considered.

Japanese words, like *"ha", "o"* and "*deha*" which have no correspondent in the French language, sometimes provoke errors. For example, in the sixth input sentence, *"kodomo deha ni"* is resulted instead of *"kodomo ni"*. In that case, *"deha"* was assumed to depend on "*kodomo"* and follow it. As far as the correct translation rate is observed, that dependency is true in most of cases. Further study on emplacement of words having no correspondent in the source sentence is still necessary.

The presence of French words in the translation result manifests the lack of resources. In the second sentence, "**tasse**" and "**remplir**" were not translated. In the first sentence, "**avec peine**", which would be translated *"karouzite"* was split because "**avec peine**", as a set, is not registered in the translation examples. The problem of unregistered words is assumed to be resolved as the number of translation examples increases.

When an infrequent word is a part of the common segment, it becomes less probable to discover sentence which matches structurally with the input sentence. This is because sentences containing that word are very limited. Giving such sentence a higher priority during the generation of the translation result will probably cause errors. We are also planning

Table 2: Results of the experiments

Input sentences	Length of input sentences	Correct translation rate		Difference between Exp.1 and Exp.2	Ratio of use of new examples
		2,500 examples only (Exp.1)	Corpus incremented (Exp.2)		
1-50	7.5	70.0%	72.0%	2.0%	1.1%
51-100	7.7	70.0%	74.0%	4.0%	2.4%
101-150	7.2	72.0%	72.0%	0.0%	2.4%
151-200	8.1	60.0%	66.0%	6.0%	5.3%
201-250	8.1	60.0%	68.0%	8.0%	6.0%
251-300	8.0	60.0%	70.0%	10.0%	13.6%
301-350	7.8	56.0%	64.0%	8.0%	15.4%
351-400	9.5	48.0%	62.0%	14.0%	24.2%
Average	8.0	62.0%	68.5%		

Table 3: Sample translation results

1	Input	**il danse avec peine.**
		(he dances with difficulty.)
	Result	*kare ha kanasi irete odorimasu.*
	Correct	*kare ha karouzite odorimasu.*
2	Input	**voulez-vous remplir la tasse d' eau chaude ?**
		(would you fill the cup with hot water?)
	Result	**(tasse)** *oyu ga* **(remplir)** *kudasai.*
	Correct	*chawan ni oyu o ippai irete kudasai.*
3	Input	**il m' a écrit qu' il avait neigé la veille.**
		(he wrote me that it had snowed the previous day.)
	Result	*zenjitu* **(avait neigé)** *to, kare ha kaita.*
	Correct	*zenjitu ha yuki ga hutta to, kare ha kaita.*
4	Input	**demain soir, madame S donnera un bal chez elle.**
		(tomorrow evening, Mrs. S will give a dance party at her home.)
	Result	*asita yoru* **(bal)** *S* **(donnera)** .
	Correct	*myouban S huzin ga zitaku de dansu pa-ti o mouyousimasu.*
5	Input	**laissez les moustaches comme elles sont.**
		(leave the mustaches as they are.)
	Result	*kutihige o nokosite kudasai kanozora you na.*
	Correct	*hige ha sono mama ni site oite kudasai.*
6	Input	**cette vue a fait peur à l' enfant.**
		(this scene made the child scared.)
	Result	*kono nagame ha kodomo deha ni kowagatte saseta.*
	Correct	*kono arisama ha sono kodomo o obie saseta.*
7	Input	**je vais à la gare pour prendre le rapide de 10 heures.**
		(I go to the station to catch the rapid-train at 10 a.m.)
	Result	**(rapide)** *dekite eki ni ikimasu 10 zi.*
	Correct	*10 zi no tokyuu ni noru tame ni eki ni iku tokoro desu.*
8	Input	**au revoir, monsieur, et bon voyage !**
		(good bye, sir, and have a nice journey !)
	Result	*sayounara kimi o, yoi go ryokou wo.*
	Correct	*sayounara, yoi tabi wo.*
9	Input	**nous allions par les routes et les chemins.**
		(we went over the roads and the paths.)
	Result	**(routes)** *o to* **(chemins)** *ikimasenka.*
	Correct	*watasitati ha ooki na miti ya tiisai miti o tootte ikimasita.*
10	Input	**l' homme et les animaux ont cinq sens.**
		(Man and animals have five sense organs.)
	Result	*otoko ha to* **(animaux)** *ha* **(cinq sens)** *arimasu.*
	Correct	*hito to doubutu ha itutu no kankaku o motte imasu.*

on doing further study on how to deal efficiently with infrequent words.

7 Conclusion

We have described an example-based machine translation method which is based on a division recursive of the input sentence. By dividing the sentence, using POS tags of words and links between source and target sentences, translation of sentences which are not fully covered by the matching sentence becomes possible. Long sentences can also be translated and any possible mismatch between source sentence and the similar sentence can be prevented from affecting the translation result. The method is designed especially for languages whose dictionaries or syntactic analyzers are not reliable or hardly available. On the other hand, with the link prediction method, the system can use new examples.

During the experiments, with 2,500 initial link-included corpus, 62.0% of correct translation rate was earned. Despite the small number of translation examples and the presence of sentences not following grammar rules in the spoken language, it is considered to be very promising results. On the other hand, by the comparison of the case where new translation examples were appended into the corpus and the other case, the rising trend of the difference between correct rates in both cases confirms the effectiveness of the link prediction method in the translation system.

Failures and errors are from a slight imperfection of the prediction method, a possible inapropriateness of the priority order of the selected examples, and a wrong positioning of words having no correspondent. These points will be 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 Française" for providing the EBTI tagging program for the French language.

References

[1] Hutchins J. and Somers H. (1992). "An Introduction to Machine Translation". Academic Press, London.

[2] Nagao M. (1984). "A Framework of a Mechanical Translation Between Japanese and English by Analogy Principle". Artificial and Human Intelligence, pp. 173-180.

[3] Sato S. and Nagao M. (1990). "Towards Memory-Based Machine Translation". Proceedings of COLING-90, pp. 247-252.

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

[5] Watanabe H. (1995). "A Model of Bi-Directional Transfer Mechanism Using Rule Combinations". Journal of Machine Translation. Vol. 10, No. 4, pp. 269-291.

[6] Jones D. (1996). "Analogical Natural Language Processing". UCL Press, London.

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

[8] Brill E. (1994). "Some advances in rule-based part of speech tagging". Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94). Seattle.

[9] Andriamanankasina T., Araki K., Miyanaga Y. and Tochinai K. (1997). "Machine Translation Based on the Relation between Words". Proceedings of "Towards Useful Natural Language Processing" NL Symposium.

[10] Brown P.F., Pietra S.A.D., Pietra V.J.D. and Mercer R.L. (1993). "The Mathematics of Statistical Machine Translation: Parameter estimation". Computational Linguistics. Vol. 19, No.2, pp. 263-311.

[11] Melamed D. (1997). "A Word-to-Word Model of Translational Equivalence". Proceedings of the 35th Conference of the Association for Computational Linguistics, pp. 490-497.

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

[13] Andriamanankasina T., Araki K, and Tochinai K. (1999). "Sub-Sentential Alignment Method by Analogy". Proceedings of PACLIC 13, pp. 277-284.

[14] Meguro S. (1987). "Manuel de Conversation Française". Hakusuisha.

[15] Sato F. (1990). "Locutions de base". Hakusuisha.