LEARNING TRANSLATION TEMPLATES FROM BILINGUAL TEXT

Hiroyuki KAJI, Yuuko KIDA, and Yasutsugu MORIMOTO
Systems Development Laboratory, Hitachi Ltd.
1099 Ohzenji, Asao-ku, Kawasaki 215, Japan

ABSTRACT

This paper proposes a two-phase example-based machine translation methodology which develops translation templates from examples and then translates using template matching. This method improves translation quality and facilitates customization of machine translation systems. This paper focuses on the automatic learning of translation templates. A translation template is a bilingual pair of sentences in which corresponding units (words and phrases) are coupled and replaced with variables. Correspondence between units is determined by using a bilingual dictionary and by analyzing the syntactic structure of the sentences. Syntactic ambiguity and ambiguity in correspondence between units are simultaneously resolved. All of the translation templates generated from a bilingual corpus are grouped by their source language part, and then further refined to resolve conflicts among templates whose source language parts are the same but whose target language parts are different. By using the proposed method, not only transfer rules but also knowledge for lexical selection is effectively extracted from a bilingual corpus.

1. Introduction

In the field of machine translation, there is growing interest in example-based approaches. The basic idea of example-based machine translation is to perform translation by imitating translation examples of similar sentences.[Nagao84] This is similar to a method often used by human translators. If appropriate examples are available, high-quality translations can be produced.

We are developing a two-phase example-based machine translation system which is composed of two subsystems: learning of translation templates from examples and translation based on template matching. This paper discusses in particular how to learn translation templates from examples. While most previous research in this area has focused on other aspects,[Sato90][Sumita91] we believe that automatic learning from examples is essential for implementing practical example-based machine translation systems.

One of the key issues in automatic learning is how to couple corresponding units (words and phrases) between bilingual texts. As far as we know, research done at BSO is the only work which has tackled this problem.[Sadler90] To what degree this procedure can be automated, however, has not been made clear. We have independently developed an algorithm for coupling corresponding units in bilingual texts.

This paper does not deal with the sentence aligning problem for bilingual texts,[Brown91][Gale91] although this is important for automatic learning from translation examples. Rather, it discusses an algorithm for learning translation templates which assumes that a technique for parallel sentence alignment is available.

Section 2 will present a rough sketch of our two-phase example-based machine translation system. Sections 3, 4, and 5 will then describe the details of the algorithm for learning translation templates from translation examples. And finally Section 6 will discuss the features of the proposed system.

2. Two-Phase Example-based Machine Translation

Figure 1 outlines our two-phase example-based machine

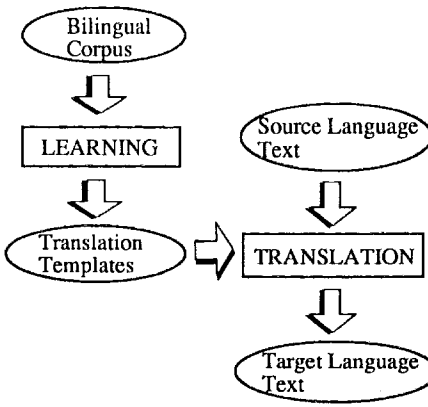

Fig.1 Two-Phase Example-Based Machine Translation

translation system. As shown in the figure, a collection of translation templates are learned from a bilingual corpus. Source language (SL) texts are translated into target language (TL) texts by using the translation templates.

Each translation template is a bilingual pair of pseudo sentences. And each pseudo sentence is a sentence which includes variables. Conditions concerning syntactic categories, semantic categories, etc. are attached to each variable. A word or phrase satisfying the conditions can be substituted for a variable. The two pseudo sentences constituting a template include the same set of variables. Parallel substitution of pairs of words or phrases, which are translations of each other, for the variables in a template produces a pair of real sentences which are translations of each other.

The learning procedure is divided into two steps. In the first step, a series of translation templates is generated from each pair of sentences in the corpus. This first step is subdivided into (a) coupling of corresponding units (words and phrases) and (b) generation of translation templates as shown in Fig. 2. The details of (a) and (b) are described in Section 3 and Section 4, respectively. In the second step, translation templates are refined to resolve conflicts among the translation templates. The details of the second step are described in Section 5.

Translation based on templates consists of (i) SL template matching, (ii) translation of words and phrases, and (iii) TL sentence generation, as shown in Fig. 3. Translation templates are regarded as directional from SL to TL, although they are actually bidirectional. First, a translation template whose SL part matches the SL sentence to be translated is retrieved. Words and phrases in the SL sentence are then bound to each variable in the template. Second, the words and phrases which are bound to variables are translated by a conventional machine translation method. And finally, a TL sentence is generated by substituting the translated words and phrases for the variables in the TL part of the translation template.

3. Coupling of Corresponding Units in Bilingual Text

An algorithm for coupling corresponding units (words and phrases) between a sentence in one language and its translation in another language is described. Although it is applicable to any pair of languages, it is explained for Japanese and English. The procedure consists of four steps: (a) analysis of Japanese sentence, (b) analysis of

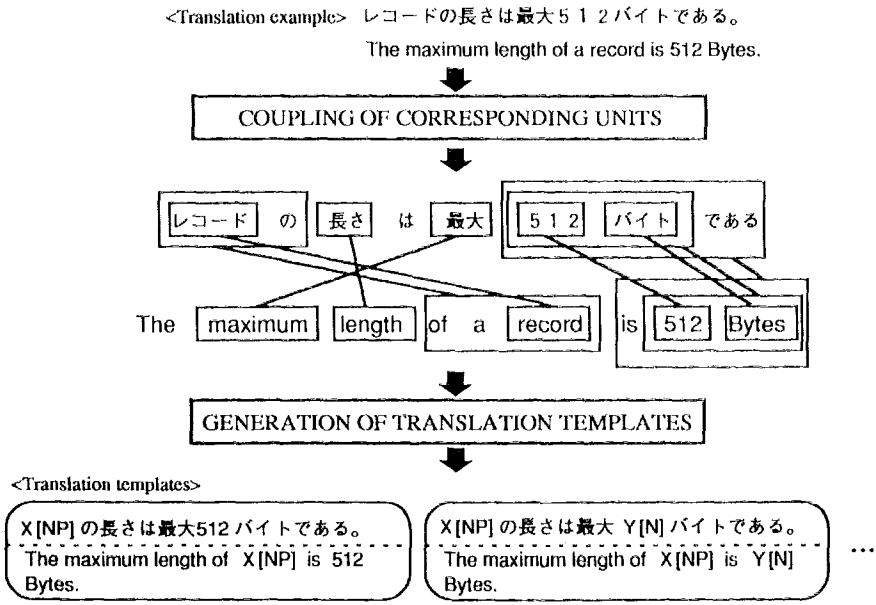

Fig.2 Generation of Translation Templates from Translation Example

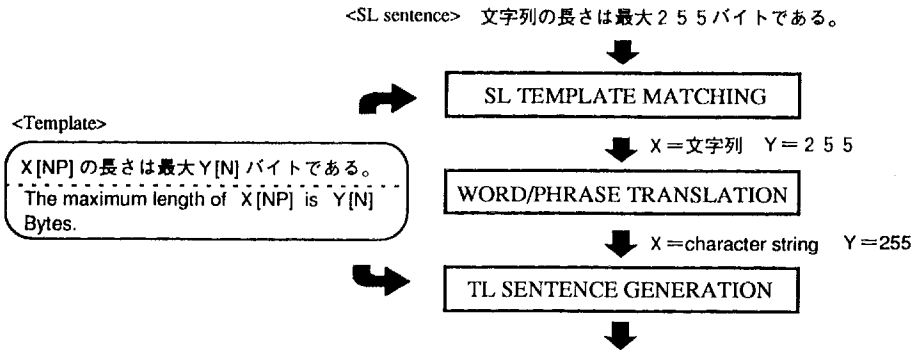

Fig.3 Translation Based on Templates

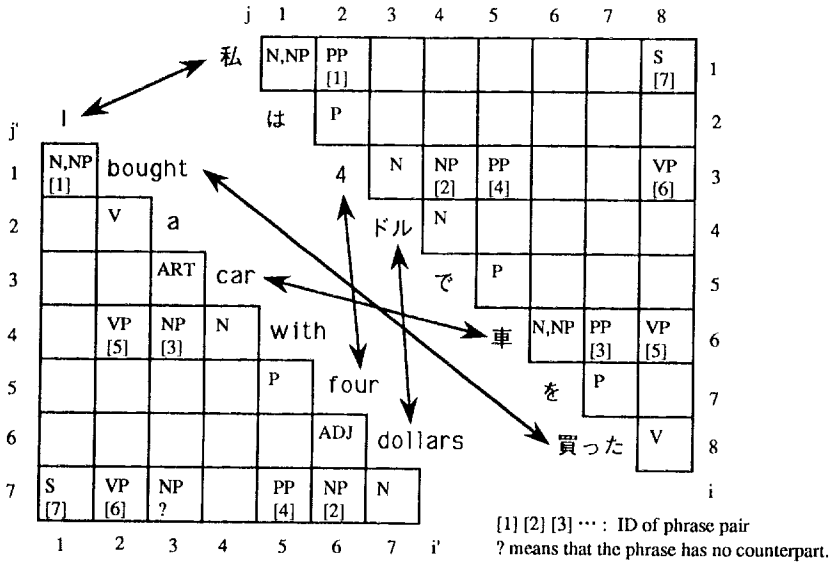

Fig.4 Sentence Analysis Tables and Coupling of Phrases

English sentence, (c) coupling of possible corresponding words between Japanese and English sentences, and (d) coupling of corresponding phrases between Japanese and English sentences.

(a) Analysis of Japanese sentence

The Japanese sentence is segmented into words by consulting a Japanese language dictionary. Then it is parsed with a parallel parsing algorithm, e.g. the CYK (Cocke-Younger-Kasami) method. As a result, a Japanese sentence analysis table is produced which expresses all possible phrases in the sentence. This Japanese sentence analysis table is a triangular matrix, as shown in the upper right portion of Fig. 4. The syntactic categories (phrase markers) of all possible phrases constituted by i-th through (i+j-1)-th words in the Japanese sentence are written in the (i,j)-element of the table. Resolution of syntactic ambiguity is postponed until the phrase coupling step.

(b) Analysis of English sentence

The English sentence is similarly analyzed and an English sentence analysis table is obtained. The English sentence analysis table is a triangular matrix, as shown in the lower left portion of Fig. 4.

(c) Coupling of possible corresponding words

Each pair of words between the Japanese sentence and its translation in English is coupled if, and only if, the pair is found in the bilingual dictionary. Obviously, there is potential ambiguity in correspondence between words if the sentence includes words which have a common translation. The most typical case is when a word occurs more than once in a sentence, as shown in Fig. 5. In this example, the correspondence between the two 'パス' and the two 'path' cannot be determined simply by consulting the bilingual dictionary. This ambiguity will therefore be resolved in the process of coupling phrases.

The coupling of words between the Japanese and English sentences is done in order to obtain candidates for variables in translation templates. We therefore restrict coupling to content words. A content word is usually replaceable with another word without affecting the grammar of the sentence. Verbs of course are closely related with sentence pattern. However, a group of verbs can produce the same sentence pattern. Therefore verbs are candidates for variables. On the other hand, function words are closely related to sentence patterns. Moreover, correspondence is not straightforward between Japanese function words and English function words. Therefore, function words should be excluded from coupling.

(d) Coupling of corresponding phrases

The Japanese and English sentence analysis tables are searched bottom up for corresponding phrases. For each phrase X in the Japanese analysis table, the English sentence analysis table is searched for a phrase Y which includes a counterpart for each word inside of X, but none for words outside of X. If a Y is found, X and Y are coupled together.

(i) Resolution of ambiguity in correspondence between words

Ambiguity in correspondence between words is resolved during the phrase coupling process as follows. Assume that a word J in the Japanese sentence has more than one counterpart in the English sentence. When a phrase X which includes J is coupled to a phrase Y in the English sentence, it is assumed that the correct counterpart for J is included in Y. This decision is highly reliable, as shorter phrases are examined before longer phrases. An example of ambiguity resolution in correspondence between words is given in Fig. 5. In this example, the ambiguity in correspondence between the two 'パス' and the two 'path' is resolved simultaneously as NP (パス 名) and NP (path name) are coupled together. Here, X (w_1 w_2 ⋯ w_n) stands for a phrase whose syntactic category is X and which is constituted by words w_1, w_2, ⋯, and w_n.

(ii) Resolution of syntactic ambiguity

A phrase X in one language sentence S is not coupled to any phrase in the other language sentence T, if T does not include a phrase which includes counterparts for all the words inside X, but none for words outside of X. This means that syntactic ambiguity is resolved implicitly in the process of coupling phrases. An example of this is shown in Fig. 4. While the English sentence analysis table contains NP (a car with four dollars), the Japanese sentence analysis table does not contain a phrase which includes '4', 'ドル', and '車' and none of the other content words. Accordingly NP (a car with four dollars) is not coupled to any phrase in the Japanese sentence. This means that NP (a car with four dollars) is rejected.

Fig. 6 shows another example of ambiguity resolution. The pair of sentences is 'A の B と C' and 'B and C of A'. While the Japanese sentence analysis table contains NP (A の B), the English sentence analysis table does not contain a phrase which includes A and B and does not include C. Accordingly NP (A の B) is rejected.

(iii) Scope of phrase

Correspondence between phrases is determined on the basis of coupled content words. There may be more than

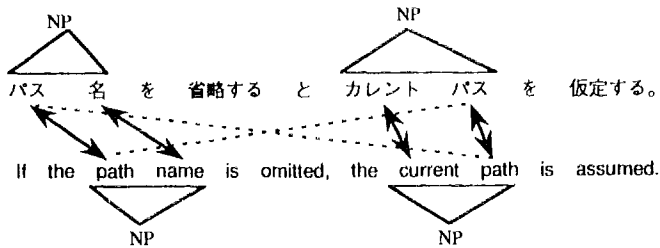

Fig.5 Resolution of Ambiguity in Correspondence between Words

one phrase containing the same set of content words. In Fig. 7(a), for example, S' (パス 名 を 省略する) and ADVP (パス 名 を 省略する と) contain the same set of content words {パス, 名, 省略する}. Likewise, S' (the path name is omitted) and ADVP (If the path name is omitted) contain the same set of content words {path, name, omit}. There are several possibilities for deciding which phrase to couple to which phrase. We decided that the smallest ones should be coupled together and the largest ones should be coupled together. In the above example, S (パス 名 を 省略する) and S' (the path name is omitted) are coupled together, and ADVP (パス 名 を省略する と) and ADVP (If the path name is omitted) are coupled together.

This strategy is also effective when a content word has no counterpart, as shown in Fig. 7(b). The bilingual dictionary does not match 'ひく' with 'play', since 'play' is not the usual translation of 'ひく'. Therefore 'ひく' has no counterpart in the sentence in Fig. 7(b). According to the strategy, however, phrases VP (ピアノ を ひく) and VP (play the piano) are coupled together.

4. Generation of Translation Templates

Each pair of coupled units is a candidate for being replaced with a variable. A template is obtained by choosing a subset of the coupled unit and assigning a unique variable to each pair in the subset. The syntactic categories (phrase markers) of the unit in the Japanese sentence are appended to the variable in the Japanese part of the template. Likewise, the syntactic categories of the unit in the English sentence are appended to the variable in the English part of the template.

The above procedure can be applied to any subset of the coupled units, as long as units which do not overlap are chosen. Accordingly, a series of translation templates can be generated from a pair of sentences. A pair of sentences and some of the translation templates generated from it are shown in Fig. 2.

A translation template need not correspond to a full sentence. Fragmentary translation templates, which correspond to fragments in a sentence, improve the flexibility of the system. The result of translation by a

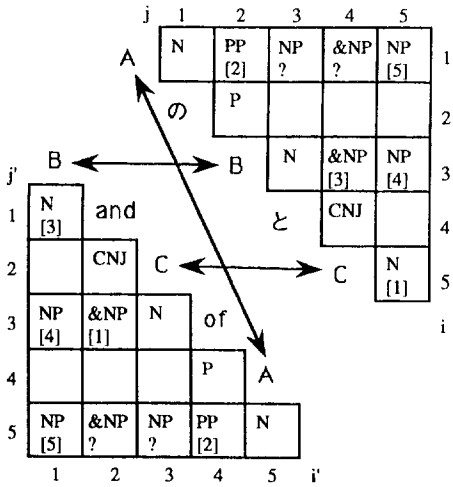

Fig.6 Resolution of Syntactic Ambiguity

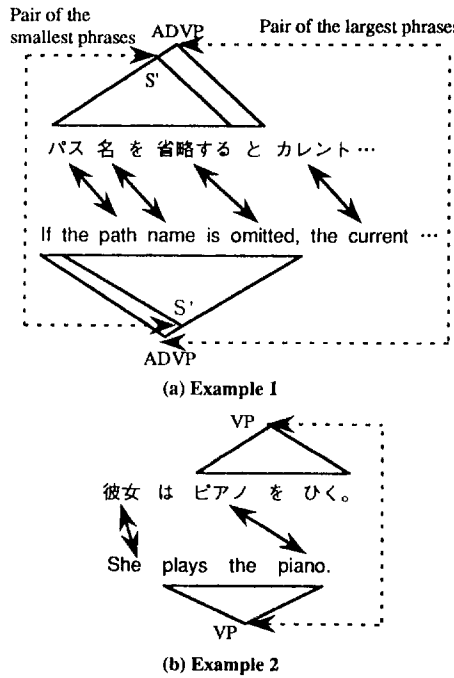

Fig.7 Coupling of Phrases and Scope of Phrase

fragmentary template may be embedded in the result of translation by another template. The fragmentary templates can also be used as a component in a conventional machine translation system.

A fragmentary translation template is generated by choosing a coupled unit pair and applying the above-described procedure to the inside of the units. The syntactic categories of the units are appended to the fragmentary translation template. An example of a fragmentary translation template is:

ADVP (X [NP] を省略すると)
/ ADVP (if X [NP] is omitted),

which is generated from the following pair of sentences.

パス名を省略するとカレントパスを仮定する。
/ If the path name is omitted, the current path is assumed.

5. Refinement of Translation Templates

Obviously the procedure described here also generates some ineffective templates, which should of course be eliminated from the collection of translation templates. The remaining ones should be refined.

In this stage, translation templates are considered to be directional. All the translation templates obtained from a bilingual corpus are grouped by their SL part, and further subgrouped by their TL part. When there is a group of templates whose SL parts are the same but whose TL parts are different, we say that they conflict with each other, because they can produce different translations for the same sentence.

If a template does not conflict with any other template, it is judged effective. It will probably produce good translations for sentences in the domain of the corpus. If a template conflicts with many templates, it is judged useless and eliminated from the collection of templates. If a template conflicts with a fewer number of templates, it is judged incomplete but possibly effective. Templates which conflict with each other are refined by examining the original translation examples from which they were generated. That is, semantic categories which distinguish each template are extracted from the original translation examples, and attached to variables in the template.

A simple example is given below. There is a conflict between templates (#1) and (#2):

(#1) play X[NP] → X[NP]をする.
(#2) play X[NP] → X[NP]をひく.

The following are translation examples from which (#1) is generated:

play baseball / 野球をする.
play tennis / テニスをする.

And the following are translation examples from which (#2) is generated:

play the piano / ピアノをひく.
play the violin / バイオリンをひく.

The conflict between (#1) and (#2) is resolved by using the semantic categories 'sport' and 'instrument' extracted from these examples. The following are the refined version of the templates:

(#1') play X[NP/sport] → X[NP] をする.
(#2') play X[NP/instrument] → X[NP] をひく.

6. Discussion

6.1 Advantages of two-phase example-based machine translation

The proposed system has the following advantages.

(1) Quality

Basically, a conventional machine translation system performs word-for-word translation. That is, a TL sentence is created from words, each of which is a TL equivalent of a word in an SL sentence. An example-based machine translation system is, in contrast, capable of creating a more flexible translation whereby elements which do not have a word-for-word correspondence are transferred as an undivided whole. We can therefore expect improvement in translation quality.

(2) Customizability

With conventional machine translation systems based on grammar rules, users are not allowed to modify the grammar rules, because they are subtly related to each other and it is difficult to assess the overall effect of rule modification. But with the example-based machine translation, users can easily customize the system for their own fields, e.g. computer manuals, by providing their own translation examples. This system is particularly suitable for a field in which similar sentences are written repeatedly.

(3) Transparency

A translation template is regarded as a transfer rule. It is easy to understand, compared to a tree-to-tree transformation rule in conventional machine translation. Translation is primarily performed by direct transfer of word string patterns. A highly transparent system can therefore be realized.

(4) Computation

Generally speaking, example-based machine translation requires large amount of computation. In the proposed architecture, however, examples are transformed beforehand into intermediate forms by extracting useful information. The amount of required computation is therefore reduced compared to a system which uses translation examples directly.

(5) Unified treatment of translation knowledge

Various kinds of knowledge for translation are extracted and represented in a single translation template framework. For example, the template in Fig. 2 is a kind of transfer rule which bridges a structural gap between Japanese and English. Lexical selection based

on co-occurrence restriction is also implemented in the framework discussed in Section 5.

6.2 Features of the algorithm for coupling corresponding units

Identifying the correspondence between units in a bilingual pair of sentences is essential for example-based machine translation. Sadler et al. have developed tools for constructing a bilingual corpus in which equivalent units are linked to each other.[Sadler90] Full automatization, however, has not yet been realized.

There are three distinguishing features of the algorithm presented in Section 3. First, the algorithm was designed on the assumption that syntactic ambiguities cannot be resolved completely by the preceding sentence analysis. Syntactic ambiguities are resolved instead in the phrase coupling process. Second, ambiguities in correspondence between words is resolved simultaneously as phrases are coupled. Third, correspondence between phrases is determined without comparing their internal structures, because structural coincidence cannot always be expected between a pair of Japanese and English sentences, even if a dependency structure is adopted. These features result in a reliable and efficient algorithm.

6.3 Is the translation template inflexible ?

The translation template may not be as flexible as the matching expression proposed by Sato.[Sato90] However, the introduction of fragmentary templates has made it sufficiently flexible.

An obvious restriction of the template is that the word order is fixed. This is inconvenient for languages, like Japanese, in which word order is flexible. However, it is not a serious problem, as the system has a learning capability. If a corpus includes sentences which differ in word order, the system will learn a set of templates which differ in word order. A more important problem to be pursued is how to deal with omissible elements. It is not easy to decide which phrases can be omitted from an example sentence. Translation templates which include descriptions of phrase omissibility, however, would certainly be effective.

7. Conclusion

We have developed an algorithm for learning translation templates from translation examples. A translation template is a bilingual pair of sentences in which corresponding units are coupled and replaced with variables. Correspondence between units is reliably identified by using a bilingual dictionary and the results of syntactic analysis of the sentences. Syntactic ambiguity and ambiguity in correspondence between units are simultaneously resolved. All translation templates generated from a bilingual corpus are grouped by their source language part, and they are then further refined to resolve conflicts among templates whose source language parts are the same but whose target language parts are different.

This algorithm makes it possible to effectively extract a variety of knowledge from a bilingual corpus. Not only is the quality of translations improved, but machine translation systems can be easily customized.

Acknowledgments

We would like to thank Mr. Shingi Domen and Dr. Fumihiko Mori for their constant support and encouragement.

References

[Brown91] Brown, P.F., et al.: "Aligning Sentences in Parallel Corpora", Proc. of 29th Annual Meeting of the ACL, pp.169-176 (June 1991).

[Gale91] Gale, W.A. and K.W. Church: "A Program for Aligning Sentences in Bilingual Corpora", Proc. of 29th Annual Meeting of the ACL, pp.177-184 (June 1991).

[Nagao84] Nagao, M.: "A Framework of a Mechanical Translation between Japanese and English by Analogy Principle", in Elithorn, A. and R. Bernerji (eds.) Artificial and Human Intelligence, North-Holland, pp.173-180 (1984).

[Sadler90] Sadler, V. and R. Vendelmans: "Pilot Implementation of a Bilingual Knowledge Bank", Proc. of COLING'90, pp.449-451 (August 1990).

[Sato90] Sato, S. and M. Nagao: "Toward Memory-based Translation", Proc. of COLING'90, pp.247-252 (August 1990).

[Sumita91] Sumita, E. and H. Iida: "Experiments and Prospects of Example-based Machine Translation", Proc. of 29th Annual Meeting of the ACL, pp.185-192 (June 1991).