Constituent Boundary Parsing for Example-Based Machine Translation

Osamu FURUSE and Hitoshi IIDA

ATR Interpreting Telecommunications Research Laboratories

Abstract

This paper proposes an effective parsing method for example-based machine translation. In this method, an input string is parsed by the top-down application of linguistic patterns consisting of variables and constituent boundaries. A constituent boundary is expressed by either a functional word or a part-of-speech bigram. When structural ambiguity occurs, the most plausible structure is selected using the total values of distance calculations in the example-based framework. Transfer-Driven Machine Translation (TDMT) achieves efficient and robust translation within the example-based framework by adopting this parsing method. Using bi-directional translation between Japanese and English, the effectiveness of this method in TDMT is also shown.

1 Introduction

Example-based frameworks are increasingly being applied to machine translation, since they can provide efficient and robust processing (Nagao, 1984; Sato, 1991; Sumita, 1992; Furuse, 1992; Watanabe, 1992). However, in order to make the best use of the advantages of an example-based framework, it is essential to effectively integrate an example-based method and source language analysis. Unfortunately, when an example-based method is combined with a source language analysis method having complex grammar rules, putting a heavy load on translation, the advantages of the example-based framework may be ruined. To achieve efficient and robust processing by the example-based framework, a lot of studies have been made for the purpose of combining source language analysis with an example-based method, and of efficiently covering the analyzed source language structure by means of transfer knowledge (Grishman, 1992; Jones, 1992; McLean, 1992; Maruyama, 1992, 1993; Nirenburg 1993).

One way to reduce the load of source language analysis is to directly apply transfer knowledge to an input string, which simultaneously executes both structural parsing and transfer knowledge application through pattern-matching. Pattern-matching does not use grammatical symbols such as "Noun Phrase", but uses surface words and non-grammatical symbols. Therefore, in pattern-matching, rule competition is reduced, and linguistic structure is expressed in a simpler manner than in grammar-based parsing. Thus, pattern-matching achieves efficient parsing. It is also useful in treating spoken language, which sometimes deviates from conventional grammar, while grammar-based parsing has difficulty treating unrestricted spoken language.

This paper proposes a constituent boundary parsing method based on pattern-matching, and shows its effectiveness for spoken language translation within the example-based framework. In our parsing method, an input string .is applied linguistic patterns expressing some linguistic constituents and their boundaries, in a top-down fashion. When structural ambiguity occurs, the most plausible structure is selected using the total values of distance calculations in the example-based framework. Since the description of a linguistic pattern is simple, it is easy to update by adding feedback.

A constituent boundary parsing method using mutual information is proposed in (Magerman 1990). This method accounts for the unrestricted natural language and is efficient. However, it tends to be inaccurate, and difficult to add feedback to, since it completely depends on statistical information without resort to a linguistic viewpoint. On the contrary, in order to achieve accurate parsing and translation, our constituent boundary parsing method implicitly incorporates grammatical information into patterns, e.g. constituent boundary description by a part-of-speech bigram, and classification of patterns according to linguistic levels such as simple sentence and noun phrase.

Transfer-Driven Machine Translation (TDMT) (Furuse, 1992, 1994) uses the constituent boundary parsing method presented in this paper, as an alternative to grammar-based analysis, and makes the best use of the example-based framework. A bidirectional translation system between Japanese and English for dialogue sentences concerning international conference registrations has been implemented (Sobashima, 1994). Experiments with the system have shown our parsing method to be effective.

Section 2 defines patterns expressed by variables and constituent boundaries. Section 3 explains a method for deriving possible English structures. Section 4 explains structural disambiguation using distance calculations in the example-based framework. Section 5 explains an example of Japanese sentence analysis using our constituent boundary parsing method, and Section 6 reports on the experimental results.

2 Pattern

A pattern represents meaningful units for linguistic structure and transfer in TDMT, and is defined as a sequence that consists of variables and symbols representing constituent boundaries. A variable corresponds to some linguistic constituent, and a constituent boundary does not allow any two variables to be adjacent. A constituent boundary is expressed by either a functional word or a part-of-speech bigram marker [1].

The explanations in this and the subsequent two sections, use English sentence parsing.

2.1 Part-of-speech

Table 1 shows the English parts-of-speech, currently used in our English-to-Japanese TDMT system. This part-of-speech system does not necessarily agree with that of conventional grammar.

Table 1 English parts-of-speech

part-of-speech	abbreviation	example
adjective	adj	*large*
adverb	adv	*exactly*
interjection	interj	*oh*
common noun	noun	*bus*
numeral	num	*eleven*
proper noun	propn	*Kyoto*
pronoun	pron	*I*
wh-word	wh	*what*
verb	verb	*go*
be-verb	be	*is*
auxiliary verb	aux	*can*
preposition	prep	*at*
conjunction	conj	*but*
determiner	det	*the*
suffix	suffix	*a.m.*

In this part-of-speech system, a be-verb, auxiliary verb, preposition, conjunction, determiner, and suffix, are classified into a functional word.

2.2 Constituent boundary marker expressed by a functional word

One problem with pattern descriptions using surface words is the necessity for a large number of patterns. To suppress the unnecessary patterns, the surface words in patterns are in principle restricted to functional words, which occur frequently, and which modify or relate content words [2].

For instance, the expression, "*go to the station*" is divided into two constituents "*go*" and "*the station*", and the preposition, "*to*" can be identified as a constituent boundary. Therefore, in parsing "*go to the station*", we use the pattern, "X *to* Y ", which has two variables X and Y, and a constituent boundary, "*to*."

2.3 Constituent boundary marker expressed by a part-of-speech bigram

The expression "*I go*" can be divided into two constituents "*I*" and "*go*." But it has no surface word that divides the expression into two constituents. In this case, a part-of-speech bigram is used as a constituent boundary.

Suppose that a constituent X is immediately followed by a constituent Y. We express a boundary-marker between X and Y by A-B, where A is a part-of-speech abbreviation of X's last word, and B is a part-of-speech abbreviation of Y's first word. For instance, "*I*" and "*go*" are a pronoun and a verb, respectively, so the marker "pron-verb" is inserted as a boundary marker into "*I go*". Namely, "*I* pron-verb *go*", i.e. with the boundary marker inserted into the original input, matches the pattern "X pron-verb Y."

2.4 Linguistic level

Patterns are classified into different linguistic levels to limit the explosion of structural ambiguity during parsing. Table 2 shows typical linguistic levels in English patterns.

Table 2 Typical levels in English patterns

level	example
beginning phrase	*excuse me but* X
compound sentence	X *when* Y
simple sentence	*I would like to* X
verb phrase	X *at* Y
noun phrase	X *of* Y, X *at* Y
compound word	X *o'clock*

[1] In this paper, variables, actual words, and part-of-speech abbreviations are expressed in capital letters, italics, and gothic, respectively.

[2] Exceptions are canned expressions such as "*I would like to*" and "*in front of*", or frequent content words such as "*what.*"

In Table 2, beginning phrase is the highest level, and compound word is the lowest. A variable on a given level is instantiated by a string described on that same level or on a lower level. For instance, in the noun phrase "X *of* Y ", the variables, X and Y cannot be instantiated by a simple sentence.

3 Derivation of Possible Structures

The algorithm for constituent boundary parsing is as follows;

(A) Assignment of morphological information to each word of an input string

(B) Insertion of constituent boundary markers

(C) Derivation of possible structures by top-down pattern matching

(D) Structural disambiguation by distance calculation

Note: we will explain (A), (B) and (C) in this section, and (D) in the next section, using the following English sentence;

(1) "*The bus leaves Kyoto at eleven a.m.*"

3.1 Assignment of morphological information

First, each word of the input string is assigned morphological information, such as its part-of-speech and conjugation form. Through this assignment, we can get the following part-of-speech sequence for (1).

(2) det, noun, verb, propn, prep, num, suffix

In addition, each word is also assigned a thesaurus code for distance calculations and an index for retrieving patterns. For instance, "*bus*" has a thesaurus code corresponding to the semantic attribute 'car.' Moreover, from the word "*at*", we can obtain the index to the pattern "X *at* Y", which is found for both verb phrase and noun phrase.

3.2 Marker insertion

A constituent boundary marker is inserted in an input string for pattern-matching. The marker is extracted from the part-of-speech sequence of an input sentence. Since such bigrams as det-noun belong to the same constituent, marker insertion by a part-of-speech bigram is restricted according to the items below.

(a) Neither A nor B is a part-of-speech relating two constituents, such as a preposition

(b) A is not a part-of-speech modifying a latter constituent, such as a determiner.

(c) B is not a part-of-speech modifying a previous constituent, such as a suffix.

We maintain a list of part-of-speech bigrams that are eligible as markers because they satisfy the above conditions. Of the bigrams in (2), "det-noun", "propn-prep", "prep-num", and "num-suffix", violate the above conditions, and are of course excluded. Thus, only "noun-verb" and "verb-propn" are inserted into sentence (1), as shown in (3).

(3) "*The bus* noun-verb *leaves* verb-propn *Kyoto at eleven a.m.*"

3.3 Application of patterns

Our pattern-matching method parses an input sentence in a top-down fashion. The highest level patterns of the input sentence are applied first; then patterns at lower levels are applied. The application procedure is as follows.

(I) Get indices to patterns from each word of the sentence. With these indices, patterns are retrieved and checked to determine if each of them can match the sentence. Then execute (II).

(II)Try to apply the highest-level patterns first. If there is a pattern that can be applied, execute (III) with respect to the variable bindings. Otherwise, execute (IV).

(III) Try to apply surface words (content words registered in a dictionary). If the application succeeds, the application for that portion is finished successfully. Otherwise, execute (II).

(IV) If the pattern to be applied is at the lowest level, the application fails. Otherwise, lower the level of the patterns and execute (II).

If pattern application finishes successfully for all portions of an input sentence, one or more source structures are obtained: since there is a possibility that more than one pattern can be applied to an expression in step (II), structural ambiguity may occur. We seek all possible structures by breadth-first application, and select the most plausible structure by the total distance value (See Section 4.4).

In step (I), indices to possible patterns are obtained from several words and bigrams in the marker-inserted sentence (3), as shown in Table 3.

Table 3 Retrieved patterns from (3)

word	retrieved pattern (linguistic level)
the	*the* X (compound word)
noun-verb	X noun-verb Y (simple sentence)
verb-propn	X verb-propn Y (verb phrase)
at	X *at* Y (verb phrase, noun phrase)
a.m.	X *a.m.* (compound word)

After step (I) is finished, steps (II)-(IV) are repeated recursively. First, the highest level pattern of the input sentence is applied. This is "X noun-verb Y ", which is defined at the simple sentence level. Next, an attempt is made to apply patterns to the variable bindings "*the bus*" and "*leaves* verb-propn *Kyoto at eleven a.m.*", which are bound to variables X and Y, respectively. To "*the bus*", at compound word level pattern "*the* X " is applied first, and the surface word "*bus*" is applied to parse "*the bus.*" Likewise, patterns and surface words are applied to the remaining part, and the application is finished successfully.

The pattern "X *at* Y " is found for both verb phrase and noun phrase. "*leaves* verb-propn *Kyoto at eleven a.m.*" thus has two possible structures, by the application of "X *at* Y." "X verb-propn Y " at the verb phrase level and "X *a.m.*" at compound word level, are also applied. Fig. 1 is the tree representation derived from the structure for sentence (1) where "X *at* Y " is a verb phrase, while Fig. 2 is a tree representation derived from the structure in which "X *at* Y " is a noun phrase. A boldface denotes the head part in each pattern. This information is utilized for extracting an input for distance calculations (See section 4.3).

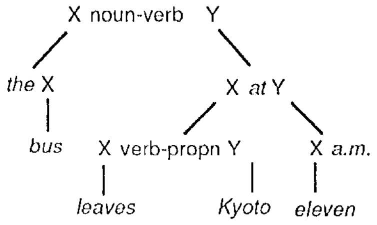

Fig. 1 Structure in which "X *at* Y " is a verb phrase

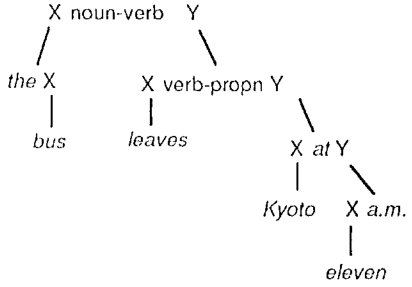

Fig. 2 Structure in which "X *at* Y " is a noun phrase

4 Distance Calculation

In this section, a method for structural disambiguation utilizing distance calculation, is described.

4.1 Distance

The distance between two words is reduced to the distance between their respective semantic attributes in a thesaurus. Words have associated thesaurus codes, which correspond to particular semantic attributes. The distance between the semantic attributes is determined according to the relationship of their positions in the hierarchy of the thesaurus, and varies from 0 to 1. The value 0 indicates that two semantic attributes belong to exactly the same category, and 1 indicates that they are unrelated.

An expression consists of words. The distance between expressions is the sum of the distance between words multiplied by each weight.

The distance is calculated quickly because of the simple mechanism employed. (Sumita, 1992) and (Furuse, 1992, 1994) give a detailed account of the distance calculation mechanism we are adopting.

4.2 Best-match by distance calculation

The advantages of an example-based framework are mainly due to the distance calculation, which achieves the best-match operation between the input and provided examples.

In TDMT, translation is performed by applying stored empirical transfer knowledge. In TDMT transfer knowledge, each source pattern has example words of variables and possible target patterns. The most appropriate target pattern is selected according to the calculated distance between the input words and the example words. The English pattern "X *at* Y " at the verb phrase level, corresponds to several possible

Japanese expressions, as shown in the following English-to-Japanese transfer knowledge:

X *at* Y	=> Y' *de* X'	((*present*, *conference*)..),
	Y' *ni* X'	((*stay*, *hotel*)..),
	Y' *wo* X'	((*look*, *it*)..)

The first possible target pattern is " Y' *de* X' ", with example set ((*present*, *conference*)..). We will see that this target pattern is likely to be selected to the extent that the input variable bindings are semantically similar to the example elements "*present*" and "*conference*." Within this pattern, X' is the target word corresponding to X, the result of transfer. "*present*" and "*conference*" are sample bindings for " X *at* Y ", where X = "*present*", and Y = "*conference*". The above transfer knowledge is compiled from such translation examples as the source-target pair of " *present a paper at the conference*" and "*kaigi de ronbun wo happyou-suru*", where "*kaigi*" means "*conference*" and "*happyou-suru*" means "present".

The semantic distance from the input is calculated for all examples. Then the example with the least distance from the input is chosen, and the target expression of that example is extracted. If the input is closest to (*stay*, *hotel*), "Y' *ni* X' " is chosen as the target expression.

The enrichment of examples increases the accuracy of determining the target expression and structure because conditions become more detailed.

4.3 Input of distance calculation

An input for distance calculation consists of head words in variable parts. In "X *at* Y " for the structure in Fig. 1, X and Y are substituted for the compound expressions, "*leaves* verb-propn *Kyoto*" and "*eleven a.m.*", respectively. In such cases, it is necessary to extract head words as the input for the distance calculation about "X *at* Y ".

In order to get head words, the head part is designated in each pattern (boldface in Figs. 1 and 2). For instance, the pattern "X verb-propn Y " contains the information that X is a head part. So the head of "*leaves* verb-propn *Kyoto*" is "*leaves*", and the head of "X *a.m.*" is "*a.m.*". Thus, in "X *at* Y " for the structure in Fig. 1, the input of the distance calculation is (*leaves*, *a.m.*).

Table 4 shows the result of distance calculation in "X *at* Y " in Fig. 1. The most plausible target structure "Y' *ni* X' " and its distance value 0.17 are obtained by the distance calculation.

Head words are passed upward from lower patterns to higher patterns. Since the head of the verb phrase pattern, "X *at* Y " is assigned to X, the head of "*leaves* verb-propn *Kyoto at eleven a.m.*" is "*leaves*", which is the head of "*leaves* verb-propn *Kyoto*". The head of "*the bus*" is "*bus*" from the head information that the head of "*the* X " is X. Thus, the input of the distance calculation of "X noun-verb Y " is (*bus*, *leave*).

Table 4 Result of distance calculation in "X *at* Y " in Fig. 1

input:(*leave*, *a.m.*)

target expression	closest example and its value [3]	
Y' *de* X'	(*arrive*, *a.m.*)	0.17
Y' *ni* X'	(*serve*, *reception*)	0.67
Y' *wo* X'	(*look*, *it*)	1.00

4.4 Structural disambiguation

Distance calculation selects not only the most plausible target expression but also the most plausible source structure. When structural ambiguity occurs, the most appropriate structure is selected by computing the totals for all possible combinations of partial distance values. The structure with the least total distance is judged most consistent with empirical knowledge, and is chosen as the most plausible structure (Furuse 1992, 1994; Sumita 1993).

Table 5 shows the result of each partial distance calculation for the structure in Fig. 1. From Table 5, we get the total distance value 1.17 for the structure in Fig. 1.

Table 5 Result of each partial distance calculation for the structure in Fig. 1

source	chosen target	distance value
the X	X'	0.33
X noun-verb Y	X' *wa* Y'	0.67
X verb-propn Y	Y' *wo* X'	0.00
X *at* Y	Y'' *ni* X'	0.17
X *a.m.*	*gozen* X' *ji*	0.00

The difference in total distance value between two possible structures for sentence (1) is due only to the distance value of "X *at* Y ", for the structure in Figs. 1 and 2. For the structure in Fig. 2, the distance value of "X *at* Y " at the noun phrase level is given as 0.83, as shown in Table 6, and is given a total distance of 1.83. Thus, the structure in Fig. 1 is selected as the

[3] These values were computed based on the present transfer knowledge of the TDMT system.

appropriate result because it has the least total distance value.

Table 6 Result of distance calculation in "X *at* Y " in Fig. 2

input:(*Kyoto, a.m.*)

target expression	closest example and its value	
Y' *no* X'	(*room, hotel*)	0.83
Y' *deno* X'	(*language, conference*)	1.00

In machine translation, it is important to disambiguate the possible structures, because a difference in structure may bring about a translation difference. For instance, the structures in Figs.1 and 2 give different Japanese translations (4) and (5), respectively. (4) is selected because it is generated from the best structure with the least total distance value.

(4) *basu wa gozen 11 ji* <u>*ni*</u> *Kyoto wo de masu*[4]

(5) *basu wa gozen 11 ji* <u>*no*</u> *Kyoto wo de masu*

5 Constituent Boundary Parsing in Japanese

Since a postposition is quite often used as a case-particle in Japanese, the boundary markers expressed by a part-of-speech bigram may not be used less frequently than in English. However, in spoken Japanese, postpositions are frequently omitted. The Japanese sentence "*Kochira wa jimukyoku*" where *kochira* means *this* and *jimukyoku* means "*office*", is translated into the English sentence "*This is the office*" by applying transfer knowledge such as the following[5]:

X *wa* Y => X' *be* Y'

But postpositions are often omitted in natural spoken Japanese, e.g. in the sentence "*Kochira jimukyoku.*" The sentence can thus be divided into two noun phrases, "*kochira*" and "*jimukyoku.*" "*kochira*" is a pronoun, and "*jimukyoku*" is a noun. So, using the bigram method of marking boundaries, we get "*Kochira* pron-noun *jimukyoku*", where the bigram "pron-noun" was inserted. The English sentence "*This is the office*" can then be produced by applying the following transfer knowledge for the pattern "X pron-noun Y ";

X pron-noun Y => X' *be* Y'

In Japanese adnominal expressions, too, constituent boundary markers are inserted between the modifier and the modified.

6 Results

We have evaluated the efficiency of our parsing method by utilizing a Japanese-to-English (JE) and English-to-Japanese (EJ) TDMT prototype system (Furuse 1994; Sobashima 1994), which is running on a Symbolics XL1200, a LISP machine with 10MIPS performance. The system's domain is inquiries concerning international conference registrations. The efficency is evaluated with 154 Japanese sentences and 138 corresponding English sentences, which are extracted from 10 dialogues in the domain. The system has about 500 source patterns for JE translation and about 350 source patterns for EJ translation.

The test sentences mentioned above have already been trained to investigate the efficiency of the method, and can be parsed correctly by the system. Table 7 outlines the 154 Japanese sentences and 138 corresponding English sentences.

Table 7 Outline of test sentences

	Japanese	English
words per input sentence	9.8	8.7
average number of possible structures	1.5	4.8

An English sentence tends to have more structural ambiguities than a Japanese sentence, because of PP-attachment, the phenomenon that an English preposition produces both a noun verb phrase and a noun phase. In contrast, the Japanese postposition does not generally produce different-level constituents.

Table 8 shows how much time it takes to reach the best structure and translation output in our JE and EJ TDMT system. The processing time for distance calculation includes structural disambiguation in addition to target pattern selection.

This demonstrates that the our parsing method can get the best structure and translation output quickly within the example-based framework.

[4] "*basu*", "*de*", and "*masu*" mean "*bus*", "*leave*", and a polite sentence-final form, respectively.

[5] For simplicity, examples and other possible target expressions are omitted.

Table 8 Processing time for the TDMT system

	JE	EJ [6]
derivation of possible structures	0.25 (sec)	0.17
distance calculation	1.32	0.14
whole translation	2.17	1.07

7 Concluding Remarks

A constituent boundary parsing method for example-based machine translation has been proposed. Linguistic patterns consisting of variables and constituent boundaries, are applied to an input string in a top-down fashion, and the possible structures can be disambigutated using distance calculation by the example-based framework. This method is efficient, and useful for parsing both Japanese and English sentences. The TDMT system, which bidirectionally translates between Japanese and English within the example-based framework, utilizes this parsing method and achieves efficient and robust spoken language translation.

By introducing linguistic information to more patterns, there is a possibility that this method can also be utilized for rule-based MT, deep semantic analysis, and so on. We will improve our parser by increasing the number of training sentences, and test its accuracy on open data.

Acknowledgements

The authors would like to thank the members of ATR Interpreting Telecommunications Research Laboratories for their comments on various parts of this research. Special thanks are due to Kohei Habara and Yasuhiro Yamazaki, for their support of this research.

Bibliography

Furuse, O., and Iida, H. (1992). Cooperation between Transfer and Analysis in Example-Based Framework. Proc. of COLING-92, pp.645-651.

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

Grishman, R., and Kosaka, M. (1992). Combining Rationalist and Empiricist Approaches to Machine Translatioin. Proc. of TMI-92, pp.263-274.

Jones, D. (1992). Non-hybrid Example-based Machine Translation Architectures. Proc. of TMI-92, pp.163-171.

McLean, I. J. (1992). Example-Based Machine Translation using Connectionist Matching. Proc. of TMI-92, pp.35-43.

Margerman, D. M., and Marcus, M. P. (1990). Parsing a Natural Language Using Mutual Information Statistics. Proc. of AAAI 90, pp.984-989.

Maruyama, H., and Watanabe, H. (1992). Tree Cover Search Algorithm for Example-Based Translation. Proc. of TMI-92, pp.173-184.

Maruyama, H. (1993). Pattern-Based Translation: Context-Free Transducer and Its Application to Practical NLP. Proc. of Natural Language Processing Pacific Rim Symposium '93, pp.232-237.

Nagao, M. (1984). A framework of a mechanical translation between Japanese and English by analogy principle. in Artificial and Human Intelligence, eds. Elithorn, A. and Banerji, R., North-Holland , pp.173-180.

Nirenburg, S., Domashnev, C., and Grannes, D.J. (1993). Two Approaches to Matching in Example-Based Machine Translation. Proc. of TMI-93, pp.47-57.

Sato S. (1991). Example-Based Machine Translation. Doctorial Thesis, Kyoto University.

Sobashima, Y., Furuse, O., Akamine, S., Kawai, J., and Iida, H. (1994). A Bidirectional Trnasfer-Driven Machine Translation System for Spoken Dialogues. Proc. of COLING-94.

Sumita, E. and Iida, H. (1992). Example-Based Transfer of Japanese Adnominal Particles into English. IEICE TRANS. INF. & SYST., Vol.E75-D, No.4, pp.585-594.

Sumita, E., Furuse, O.,and Iida, H. (1993). An Example-Based Disambiguation of Prepositional Phrase Attachment. Proc. of TMI-93, pp.80-91.

Watanabe, H. (1992). Similarity-Driven Transfer System. Proc. of COLING-92, pp.770-776.

[6] The distance calculation time in EJ translation is short, since the system has not yet learned enough translation examples concerning EJ translation.