Pattern Dictionary Development based on Non-Compositional Language Model for Japanese Compound and Complex Sentences

Satoru Ikehara[1], Masato Tokuhisa[1], Jin'ichi Murakami[1],
Masashi Saraki[2], Masahiro Miyazaki[3] and Naoshi Ikeda[4],

[1] Tottori University, Tottori-city, 680-8552 Japan.
{ikehara, tokuhisa, murakami}@ike.tottori-u.ac.jp
[2] Nihon University, Tokyo, 101-0061 Japan. saraki@st.rim.or.jp
[3] Niigata University, Niigata-city, 950-2102 Japan. miyazaki@ie.niigata-u.ac.jp
[4] Gifu University, Gifu-city, 501-1112 Japan. ikeda@info.gifu-u.ac.jp

Abstract. A large-scale sentence pattern dictionary (SP-dictionary) for Japanese compound and complex sentences has been developed. The dictionary has been compiled based on the *non-compositional language model*. Sentences with 2 or 3 predicates are extracted from a Japanese-to-English parallel corpus of 1 million sentences, and the compositional constituents contained within them are generalized to produce a SP-dictionary containing a total of 215,000 pattern pairs. In evaluation tests, the SP-dictionary achieved a syntactic coverage of 92% and a semantic coverage of 70%.

Key Words: Pattern Dictionary, Machine Translation, Language Model

1. Introduction

A wide variety of MT methods are being studied [1, 2, 3], including *pattern-based MT* [4, 5], *transfer methods*, and *example-based MT* [6, 7, 8], but it is proving to be difficult to obtain high-quality translations for disparate language groups such as English and Japanese. *Statistical MT* have been attracting some interest recently [9, 10, 11], but it is not easy to improve the quality of translations. Most practical systems still employ the *transfer method*, which is based on *compositional semantics*. A problem with this method is that it produces translations by separating the syntactic structure from the semantics and is thus liable to lose the meaning of the source text.

Better translation quality can be expected from *pattern-based MT* where the syntactic structure and semantics are handled together. However, this method requires immense pattern dictionaries which are difficult to develop, and so far this method has only been employed in hybrid systems [12, 13] where small-scale pattern dictionaries for specific fields are used to supplement a conventional *transfer method.*

Example-based MT has been expected to resolve this problem. This method obtains translations by substituting semantically similar elements in structurally matching translation examples, hence there is no need to prepare

a pattern dictionary. However, the substitutable elements depend on translation examples. This made it impossible to judge them at real time. This problem could be addressed by manually tagging each example beforehand, but the resulting method would be just another *pattern-based MT.*

This problem [14] has been partially resolved by a highly comprehensive valency pattern dictionary called *Goi Taikei* (A-Japanese-Lexicon) [15]. This dictionary contains 17,000 pattern pairs for the semantic analysis in the Japanese-to-English MT system ALT-J/E [16]. High quality translations with the accuracy of more than 90% has been performed for simple Japanese sentences, but there are still cases where a suitable translated sentence structure cannot necessarily be obtained. A valency pattern expresses the semantic relationship between independent words. The meaning of subordinate words (particles, auxiliary verbs, etc.) is dealt with separately, hence the original meaning is sometimes lost. Addressing this problem requires a mechanism that deals with the meaning of subordinate words within the sentence structure as a whole.

In order to realize such a mechanism, we propose a language model that focuses on the non-compositional expressions, and a method for creating patterns based on this model. This method obtains pattern pairs from parallel corpus by the semi-automatic generalization of compositional constituents.

2. Non-Compositional Language Model

2. 1 Compositional constituents and non-compositional constituents

In the framework of expressions that come to mind during the process where a speaker is forming a concept, there are two types of constituents to consider. One is those that cause the overall meaning to be lost when they are substituted with other alternative constituents. And the other is those that do not cause the overall meaning to be lost. The former are referred to as *N-constituents (Non-compositional constituents)*, and the latter are referred to as *C-constituents (Compositional-constituents).*

Definition 1: C- constituents and *N-constituents*
C-constituent is defined as a constituent which is interchangeable with other constituents without changing *the meaning of an expression structure.* All other constituents are *N-constituents*.

Definition 2: C-expressions and *N-expression*
C-expression (Compositional expression) is defined as an expression consisting of *C-constituents*, and *N-expression (Non-compositional expression)* is defined as an expression comprising one or more *N-constituents*.

Where a *constituent* is a part of an *expression* consisting of one or more words, one *constituent* can constitute one *expression*.

Before applying these definitions to actual linguistic expressions, *the meaning of an expression structure* is needed to be defined. Although a great deal of research has been made concerning the meaning of linguistic expressions, any statement is nothing more than a symbol as far as processing by a computer is concerned, and hence we just need to express meanings in a system of symbols that is free from semantic inconsistencies. In this study, considering applications to Japanese-to-English MT, *the meaning of expression structures* is defined in terms of an English expression.

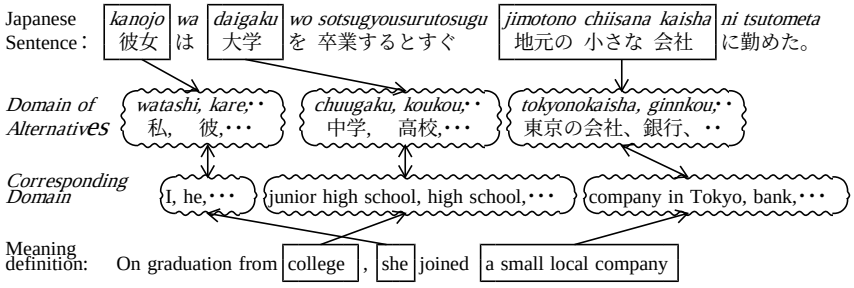

Fig. 1 Example of C-constituents

In Figure 1, the source sentence is a Japanese expression expressing a relationship between two events. *The meaning of the expression structure* is *Immediately after performing one action, somebody performed the other action.* This meaning is defined by using the English expression. For the constituent such as 彼女*(she),* 大学*(college)* and 地元の小さな会社 *(small local company),* there is a domain of substitutable constituents that doesn't change *the meaning of the expression structure*, therefore these are C-constituents.

2. 2 Characteristics of C-*constituents*

From the above definitions, it can be pointed out that a C-constituent possesses the following four important characteristics. From these characteristics, it is possible to obtain important guidelines for pattern-forming.

(1) Language pair dependence of C-constituent

Since one linguistic expression is used to define the meaning of another, the number and scope of C-constituents depends on the language pair. For languages that belong to the same group, the scope of C-constituents is large, while for disparate language groups it is expected to be smaller, as reflected in the different levels of difficulty of translating between the languages.

(2) Finite choice for alternative constituents

Although C-constituents can be substituted, that does not mean they can be substituted with anything at all. The range that can be substituted is limited both grammatically and semantically, thus this must be indicated in the pat-

tern as the "domain" of the constituent.

(3) C-constituent dependent on constituent selection

The scope of constituents is determined arbitrarily. Hence whether a constituent is compositional or non-compositional depends on how the constituent is chosen. Accordingly, to obtain general-purpose patterns, it is better to increase the number of C-constituents.

(4) Simultaneity of a C-constituent and an N-expression

A so-called C-constituent is only compositional when seen in the context of the entire expression, and itself may actually be a N-expression.

2. 3 Language Model

According to definition 1, a linguistic expression consists of C-constituents and N-constituents. According to characteristic (3), if we select a C-constituent from an expression with a meaningful range (e.g., word, phrase or clause), a C-constituent may itself also be an N-expression according to characteristic (4). Consequently a linguistic expression can generally be expressed with the language model shown in Fig. 2.

As this figure shows, when C-constituents are repeatedly extracted from N-expressions, the end result is an N-expressions that contains no C-constituents. Although the resulting N-expression may just be a single word, it could also be an idiomatic phrase that has no substitutable constituents. Thus, in this language model, linguistic expressions can be articulated into one or more N-expressions and zero or more N-constituents.

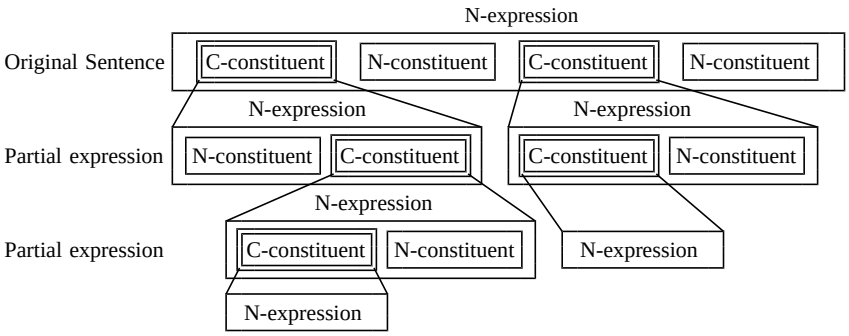

Fig. 2 Non-compositional language model

2. 4 Patterns for N-expressions

An important aspect of the language model is that the N-expressions that appear at each stage of the articulation are meaningful expression units. In this element decomposition process, loss of the original meaning can be avoided by using a semantic dictionary for N-expressions at each stage. For example, if linguistic expressions are classified into sentences, clauses and phrases, and semantic dictionaries are constructed for N-expressions at each of these

levels, then this would constitute the bulk of a mechanism for assimilating the meaning of entire sentences.

It is thought that patterns are a suitable framework for expressing the syntactic structure of N-expressions, because:

(a) a N-constituent cannot be substituted with another constituent, thus a literal description is appropriate, and
(b) the order in which C- and N-constituents appear is often fixed, thus there is thought to be little scope for variation.

Therefore, in this study we will use a pattern-forming approach for meaningful N-expressions.

3. Development of SP (Sentence Pattern)-dictionary

According to our language model, three kind of expression patterns (compound and complex sentence patterns, simple sentence patterns and phrase patterns) will be almost sufficient to cover Japanese expressions.

In this study, complex and compound sentences were targeted because the *Goi Taikei* [15] can gives good translations for most of simple sentences. But, complex and compound sentences are very difficult to obtain good translation results by the conventional MT systems. The number of predicates was limited to 2 or 3 because it is thought that complex and compound sentences with four or more predicates can often be interpreted by breaking them down into sentences with three predicates or fewer.

3. 1 The principles of pattern-forming

The Japanese-English parallel corpus is a typical example where the meaning of Japanese expressions is defined with English expressions. And when translation example is considered, the following two types of C-constituents can occur:

(1) cases where there is a constituent in the English expression that corresponds to a constituent in the Japanese expression, and
(2) cases where a constituent in the Japanese expression has no corresponding constituent in the English expression, but deleting this constituent from the Japanese expression does not cause any change in the corresponding English expression.

SP pairs were therefore produced by extracting components corresponding to these two cases from parallel corpus, and generalizing the results.

3. 2 SP generation procedure

First, a parallel corpus was created by collecting together a sentence pair of 1 million basic Japanese sentences. From this corpus, 150,000 translation examples for compound and complex sentences with two or three predicates were extracted. Then, using resources such as Japanese-English word dic-

tionaries, the semantic correspondence relationships between the constituents were extracted and converted into *variables*, *functions, symbols* in the following three stages to produce a SP-dictionary.

- *Word-level generalization:* compositional independent words (nouns, verbs, adjectives, etc.) are replaced by *variables*.
- *Phrase-level generalization:* compositional phrases (noun phrases, verb phrases, etc.) are replaced by *variables*.
- *Clause-level generalization:* compositional clauses (adnominal clauses and continuous clauses) are replaced by *variables*.

For C-constituents that can be semi-automatically recognized as such, the generalization is also performed semi-automatically.

3. 3 Examples of SPs

An example of a SP is shown in Table 1. The meanings of the *variables*, *functions,* etc. used in this table are shown below.

Table 1. Examples of generated SPs

word-level SP	
Japanese SP	*ha* *te* *wo* *ni* #1[N1(G4)は]/V2(R3003)て/N3(G932)を/N4(G447)に/V5(R1809).*tekita*。
English SP	[N1|I]was so AJ(V2)as to V5 #1[N1^*poss*]N3 at N4.
Example	*ukkarisite* *teikikennwo* *ieni* *wasuretekita* うっかりして 定期券を 家に 忘れてきた。 I was so careless as to leave my season ticket at home.
phrase-level SP	
Japanese SP	*ha* *ni* *nodakara* NP1(G1022)は/V2(R1513).*ta*/N3(G2449)に/V4(R9100).*teiru* のだから/N5(N1453).*dantei*。
English SP	NP1 is AJ(N5) in that it V4 on AJ(V2) N3.
Example	*sonoketsuronwa* *ayamattazenteini motozuite* *irunodakara* *ayamaridearu* その結論は 誤った前提に 基づいて いるのだから 誤りである。 The conclusion is wrong in that it is based on a false premise.
clause-level SP	
Japanese SP	*node* *niatatteha* CL1(G2492).*tearu* ので、N2(G2005)に当たっては/VP3(R3901).*gimu*
English SP	*so+that* (CL1, VP3.*must.passive* with subj (CL1)^*poss* N2)
Example	*sorewa* *kiwamete* *yuudokude arunode* *siyouniatattewa* *juunibunni* それは 極めて 有毒であるので、 使用に当たっては 十二分に *chuuisinakerebanaranai* 注意しなくてはならない。 It is significantly toxic so that great caution must be taken with its use

Word-level SPs: ① N1, N3, N4: *Noun variables.* ② V2, V5: *Verb variables.*

Here, attached bracket represents semantic attribute numbers specifying semantic constraints on a variable. ③#1[...]: Omissible constituents. ④ /: Place of a constituent that need not appear. ⑤*.tekita*: Function for specifying a predicate suffix. ⑥ AJ(V2): Adjectival form of the value of verb variable V2. ⑦ N1^poss: Value of N1 transformed into possessive case.

Phrase-level SPs ① NP1: *Noun phrase variable.*

Clause-level SPs ① CL1: *Clause variable.* ② *so+that* (..., ...): *A sentence generation function* for *so that* sentence structure. ③ subj(CL): Function that extracts the subject from the value of *a clause variable.*

3. 4 The number of different SPs

Table 2 shows the number of SPs in the resulting SP-dictionary and the number of constituents replaced by *variables* at each level of generalization.

Table 2. Number of different SPs and Ratio of C-constituents

Type of SPs	*word-level*	*phrase-level*	*clause-level*	Total
No. of pattern pairs	122,642 pairs	80,130 pairs	12,450 pairs	215,222 pairs
Ratio of C-constituents	472,521/763,968 = 62 %	102,000/463,636 = 22 %	11,486/267,601 = 4.3 %	----

In Table 2, compared to the number of SPs of *word-level* and *phrase-level* SPs, the number of *clause level* SPs was particularly small. This indicates that most of the clauses in the parallel corpus are N-constituents which are impossible to generalize. The proportion of generalized C-constituents were 62% at the *word level* and 22% at the *phrase level*, but just 4.3% at the *clause level*.

For N-constituents, a semantically suitable translated result cannot be obtained when the constituent is extracted, translated and incorporated into the original sentence. Looking at the parallel corpus, most of the English translations of Japanese compound and complex sentences are simple sentences whose structures are very diverse. Regarding the results of Table 2, in the case of Japanese-to-English MT, high-quality translations cannot be achieved by conventional MT method based on *compositional semantics*.

4. The Coverage of the SP-dictionary

4. 1 Experimental conditions

A *pattern parser* that compares input sentences against the SP-dictionary was used to evaluate the coverage of the SP-dictionary. The experiments were conducted by *cross-validation* manner and ten thousand input sentences were used. These were randomly selected from the example sentences used for creating the SPs. Since the input sentences will always match the SPs from which they were created, matches of this type were ignored and the

evaluation was restricted matches to other SPs.

An input sentence many times matches to more than one SP and not all of them are necessarily correct. Therefore, the coverage was evaluated according to the following four parameters:

- *Matched pattern ratio* (R): The ratio of input sentences that are matched to at least one SP (*syntactic coverage*)
- *Precision* (P1): The ratio of matched SPs that are semantically correct
- *Cumulative precision* (P2): The ratio of matched SPs for which there is one or more semantically correct SP
- *Semantic coverage* (C): The ratio of input sentences for which there is one or more semantically correct SP (R ~P2)

4. 2 Saturation of *matched pattern ratio*

Fig. 3 shows the relationship between the number of SPs and the *matched pattern ratio*. As you can see, there is a pronounced tendency for the *matched pattern ratio* to become saturated. When the SPs on the horizontal axis are rearranged in order of their frequency of appearance, the rate of saturation becomes about 5 times faster.

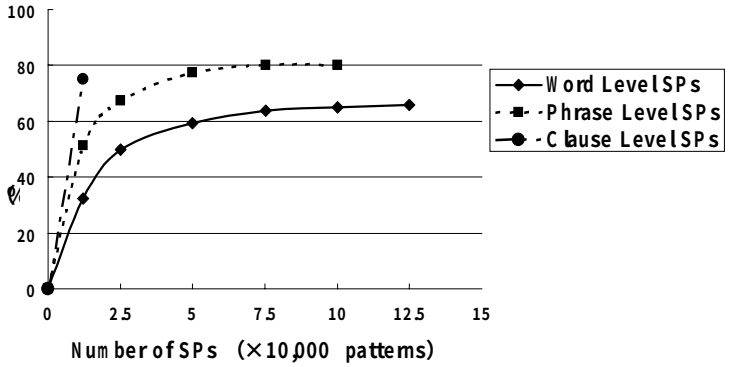

Fig. 3. Saturation of *Matched pattern ratio*

According to the previous study [17], the number of valency patterns required to more or less completely cover all simple sentences was estimated to be somewhere in the tens of thousands. We can say that the number of required SPs for complex and compound sentences is also expected to converge somewhere in the tens of thousands or thereabouts.

4. 3 *Matched pattern ratio* and *precision*

Table 3 shows the evaluation results. It was shown that 91.8% of the input sentences are covered syntactically by the whole dictionary. However, there were also many cases of matches to semantically inappropriate SPs, and the

semantic coverage decreased to 70% when these were eliminated. The number of *clause-level SPs* was just one tenth the number of *word-level SPs*, but had comparatively high coverage.

Table 3. Coverage of SP-dictionary

Type of SPs	R (*Matched pattern ratio*)	P1 (*Precision*)	P2 (*Cumulative precision*)	C=RxP2 (*Semantic coverage*)
Word Level	64.7 %	25 %	67 %	43.3 %
Phrase Level	80.0 %	29 %	69 %	55.2 %
Clause Level	73.7 %	13 %	68%	50.1 %
Total	91.8 %	——	——	70 %

4. 4 *Semantic coverage*

Since semantic ambiguity is small in the order of *word-level, phrase-level* and *clause-level SPs,* it is probably better to select and use the most semantically appropriate SP based on this sequence. Fig. 4 shows the ratio of SPs that are used when they are selected based on this sequence.

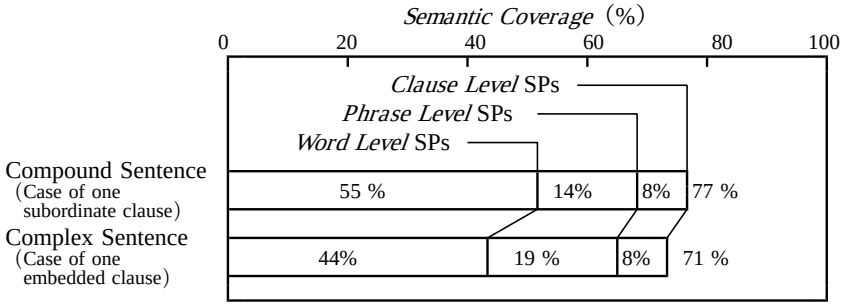

Fig. 4 *Semantic coverage* of SP-dictionary

As Fig. 4 shows, about 3/4 of the meanings of Japanese compound and complex sentences are covered by the SP-dictionary. When MT is performed using the SP-dictionary, it is estimated that word-level SPs will be used for about half of the complex and compound sentences, while phrase-level and clause-level SPs will be applied to the other half.

5. Concluding Remarks

An Non-compositional language model was proposed and, based on this model, a sentence pattern dictionary was developed for Japanese compound and complex sentences. This dictionary contains 123,000 *word-level*, 80,000 *phrase-level* and 12,000 *clause-level* sentence pattern pairs (215,000 in total).

According to the results, the compositional constituents that could be generalized were 62% for independent words, 22% for phrases, whereas only 4.3% for clauses. This result shows that in Japanese-to-English MT hardly any Japanese compound and complex sentences can be translated into English as shown in a parallel corpus when they are translated by separating them into multiple simple sentences and then recombined.

Also, in evaluation tests of a SP-dictionary, the syntactic coverage was found to be 92%, while the semantic coverage was 70%. It is therefore proved that the SP-dictionary is very promising for Japanese to English MT.

Acknowledgements

This study was performed with the support of the *Core Research for Evolutional Science and Technology* (CREST) program of the *Japan Science and Technology Agency* (JST). Our sincere thanks go out to everyone concerned and to all the research group members.

References

1. Nagao. M.: Natural Language Processing, Iwanami Publisher (1996)
2. Ikehara. S.: Machine Translation, in Information Processing for Language, Iwanami Publisher (1998) 95-148
3. Tanaka, H. (Eds): Natural Language Processing - Fundamentals and Applications, Iwanami Publisher (1998)
4. Takeda, K.: Pattern-based Machine Translation, COLING, Vol. 2 (1996) 1155-1158
5. Watanabe, H. and Takeda, K.: A Pattern-based machine translation system extended by example based processing, COLING (1998) 1369-1373
6. Nagao, M.: A Framework of a Mechanical Translation between Japanese and English by Analogy Principle, in Artificial and Human Intelligence, North-Holland (1984) 173-180
7. Sato, S.: An example based translation and system, COLING (1992) 1259-1263
8. Brown, R. D.: Adding Linguistic Knowledge to a Lexical Example-Based Translation System, TMI 99 (1999) 22-32
9. Brown, P. F., John, C. S., Pietra, D., Jelinek, F. J., Lfferty, D. , Mercar, R. L. and Roossin, P. S.: A Statistical Approach to Machine Translation, Computational Linguistics, Vol. 16, No. 2 (1990) 79-85
10. Watanabe, T. and Sumita, E.: Bi-directional Decoding for Statistical Machine Translation, COLING (2002) 1075-1085
11. Vogel, S., Zhang, Y., Huang, F., Tribble, A., Venugopal, A., Zhao, B. and Waibel, A.: The CMU statistical machine translation system. MT Summit IX (2003) 402-409
12. Jung, H., Yuh, S., Kim, T., Park, S.: A Pattern-Based Approach Using Compound Unit Recognition and Its Hybridization with Rule-Based Translation, Computational Intelligence, Vol. 15, No. 2 (1999) 114-127
13. Uchino, H., Shirai, S., Yokoo, A., Ooyama, Y. and Furuse, K.: News Flash Translation System of ALTFLASH, IEICE Transactions, Vol. J84-D-II, No. 6 (2001) 1168-117
14. Ikehara, S.: Challenges to basic problems of NLP, J. of JSAI, Vol. 16, No. 3 (2001) 522-430
15. Ikehara, S., Miyazaki, M., Shirai, S., Yokoo, A., Nakaiwa, H., Ogura, K., Ooyama, Y. and Hayashi, Y.: *Goi Taikei* (A-Japanese Lexicon), Iwanami Publisher (1997)
16. Ikehara, S., Miyazaki, M., Shirai, S. and Hayashi, Y.: Speaker's conception and multi-level MT, J. of IPSJ, Vol. 28, No. 12 (1987) 1269-1279
17. Shirai, S., Ikehara, S., Yokoo, A. and Inoue, H.: The quantity of valency pattern pairs required for Japanese to English MT and their compilation. NLPRS '95, Vol. 1, (1995) 443-448.