NARA: A Two-way Simultaneous Interpretation System between Korean and Japanese -A methodological study-

Hee Sung Chung and Tosiyasu L. Kunii

Department of Information Science
Faculty of Science, University of Tokyo
7-3-1 Hongo, Bunkyo-ku Tokyo, 113 Japan

Abstract

This paper presents a new computing model for constructing a two-way simultaneous interpretation system between Korean and Japanese. We also propose several methodological approaches to the construction of a two-way simultaneous interpretation system, and realize the two-way interpreting process as a model unifying both linguistic competence and linguistic performance. The model is verified theoretically and through actual applications.

1. Introduction

Our goal is to develop a two-way simultaneous interpretation system between Korean and Japanese. In order to achieve this goal, we have designed a specific computing model, which is a computer program based on the algorithm that formalizes the mechanism of two-way simultaneous interpretation and the correspondence of the two languages. Our computational approach consists of two parts. First, build an explicit computational model, then show the practical applicability and theoretical validity of the model. The most significant advantage of using a formal description to represent our system is in that the descriptive contents of the representative algorithm do not depend upon the conventional approaches to machine translation. We have also implemented a prototyping system NARA, a two-way simultaneous interpretation system between Korean and Japanese. In this paper, we outline the features of the system without going into the details.

2. Methodology

Our approach is intuitively motivated by Chomsky's hypothesis[1]: homogeneous communication by the same linguistic performance is possible among those who have the same linguistic competence. We take a performance theory to be the study of real time processing of languages. The performance theory cannot be developed without a competence theory. This hypothesis suggests that a key point of contact between the theory of grammar and the interpretation control is the natural link between the theory of knowledge representation and the theory of knowledge processing. That is, for two classes of languages to be interpretable by human being, there exists an interpreting procedure. Consequently, if we can show that there is an adjusted grammar for the two languages plus an adequate interpreting procedure to predetermine the mechanism of our two-way simultaneous interpretation, then we have some support for our methodology. In order to guarantee two-way simultaneous interpretation, there are several subareas to be inquired. The first is the type of representation constructed during the interpretation. The second is the method of utilizing the representation during the interpretation. The third is the measure of computational complexity during the interpretation. These three components of a complete computational model are necessary for linking the adjusted grammar with the interpretation mechanism: a representation, an algorithm, and a complexity metric. We take the following items as the subjects of methodological study.

(1) The theory of grammar

We require an adjusted grammar to be suitable for description of the two languages as input and output. It is intuitively clear that the more communicatable the adjusted grammar is, expressed by a powerful formal system, the more efficiently is the grammar interpreted. We adopt generalized phrase structure grammar(GPSG) framework[3].

(2) The notion of direct realization of interpretation

Because we need to connect competence and performance as directly as possible, one of the goals of our study is to identify rules of the grammar with the manipulative unit of interpretation in a one-to-one fashion. Thus we carefully distinguish between the grammar and the rules of interpretation. For this, we adopt the following notions as the methodological principles of our system:

1) Equivalence of grammar[5], 2) grammar cover and grammar modification[6], 3) type transparency[2], and 4) an invariant of formal languages[4].

(3) The notion of complexity measure

The direct association between unit interpretation time cost and the complexity of a sequential operation during interpretation can be measured.

3. Linguistic Data Structure and Computing Model

In order to investigate the correspondence between the two languages, we partition a grammar into independent components: segmented words, the word order, morphology, syntax, and semantics. The partition of a grammar constitutes an important step of modular decomposition into the interpretation subsystems.

3.1 Interpretation strategy of segmented word component

3.1.1 Data structure

In comparison with other symbol system, every human language has a remarkable characteristics; namely, the structure of segmented words. The utterance as a segmented word conveys a message regarding some matter, and communicates the information concerning the matter. A segmented word is a word or an ordered pair of words. Using some criteria: positional transformation, substitution and insertion, we can specify a segmented word of Korean or Japanese.

3.1.2 Word order in a segmented word of Korean or Japanese

Between Korean and Japanese, some common properties are observed, such as an agglutinative language structure and the identical word order(SOV). In addition, we sight three corresponding word order properties of segmented words between the two languages:

For some (k1, k2) ∈ Sk and (j1, j2) ∈ Sj, where Sk and Sj are a set of Korean segmented words, a set of Japanese segmented words, respectively, and I a binary relation(interpretation):

[Property 1] reflexivity

(k1,k2) <I> (j1,j2).

e.g。わが 国 <I> 우리 나라 (our nation)

[Property 2] symmetry

(j1,j2) <I> (k2,k1).

e.g.もう 一度<I> 한번 더 (one more time)

[Property 3] transitivity

(j1,no,j2) <I> (k1,k2).

e.g. 日本の 人 <I> 일본 사람 (a Japanese)

Among above properties, Property 3 depends upon Korean pragmatic information.

3.1.3 Computing model

The production form of a segmented word of Korean or Japanese can be described in the rule forms in a regular grammar, and it is right linear. Since a language L generated by some right linear grammar G is regular, there exists a finite automaton which accepts L. If L is a context-free language and s is a substitution map such that for every a ∈ V(a fixed vocabulary), s(a) is a context-free language, then s(L) is a context-free language. A type of substitution map that is of special interest is a homomorphism. If L is a regular language and h is a homomorphism, then the range of the inverse homomorphism $h^{-1}(L)$ is also regular language. And, for two given regular grammars G and G', if L(G) = L(G'), there is a sequence equivalence. Two sequences generate the same word order in the increasing length order.

3.2 Interpretation strategy of Morphological component

3.2.1. Data structure

The study of the structure of words occupies an important place within linguistics, sandwiched between phonology and syntax. Morphemes may also be partitioned into lexical and grammatical classes. Lexical morphemes are generally free, while many of the grammatical morphemes are bound.

3.2.2 Computing model

In a given Korean-Japanese (or Japanese-Korean) dictionary, let Dk be the set of morphemes of Korean, and Dj be the set of morphemes of Japanese. A mapping I between the sets is defined as follows.

$$I(Dk) = Dj$$

implying that the image of Dk is Dj; taking the inverse mapping,

$$I^{-1}(Dj) = Dk.$$

By generalizing the relation and the mapping between the two sets, we may consider the set of Korean words to be a domain, and the set of Japanese words a range. Assuming the same cardinality for both, Dk and Dj may be partitioned as shown below. Here we suppose

$$\{k1,\ k2,..kn\} \in Dk,\ \{j1,\ j2,..jm\} \in Dj.$$

(1) one-to-one $(k_i, j_i) \in Dk \times Dj$.

(2) one-to-many $(k_i, \{j_{i1}, j_{i2}, \cdots j_{in(i)}\}) \in D_k \times 2^{Dj}$

(3) many-to-many $(\{k_{i1}, k_{i2}, \cdots k_{in(i)}\}, \{j_{i1}, j_{i2}, \cdots j_{im(i)}\}) \in 2^{Dk} \times 2^{Dj}$

where, $A \times B$ is the Cartesian product of the two sets A and B, and 2^A is the a power set of a set A.

Obviously, one-to-one correspondence is isomorphic. Naturally, our attention will be focused on the one-to-many and many-to-many relations. Interpretation of these relations depends on various factors: allomorph, synonym and homonym. Thus, as for the interpretation which is dependent on synonymy or polysemy, we characterize the interpretation by specifying the canonical form, or the semantic feature instantiation, respectively.

3.3 Syntax level interpretation strategy

We examine the syntactic structure of the two languages. From the correspondence in a segmented word and word order, it is seen intuitively that they are strongly equivalent. And there is a sufficient linguistic evidence for it based on the study of experimental comparative linguistics[2]. A phrase structure preserves each lexical semantic feature of a constituent structure, and a parse tree describes the construction of syntactic representation of a sentence. Moreover, a partial tree in the whole parse tree plays a role of adjusting semantic and syntactic interpretation. Let us compare the examples of two parse tree constructions(Fig 1):

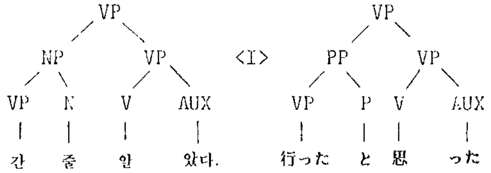


Fig 1: Syntactic trees of `(I) thought (somebody) went (to somewhere)'

It is obvious that parse trees coincide with each other in one-to-one fashion, but syntactic categories do not. This implies that two given languages, Korean and Japanese, do not generate the same set of sentential forms. Furthermore, there is no algorithm for deciding whether or not two given context-free grammars generate the same sentential forms. This is the reason why we adopt the covering grammar technique to parse the source language for interpretation.

3.4 Semantics, pragmatics and ambiguity

Semantics and pragmatics also play an important role in generating the well-formed target language. In the interpretation between Korean and Japanese, there exist several kinds of inherently ambiguous sentences which are generated only by the ambiguous grammars of both languages. (see 5.Fragments of interpretation)

4. K-J Grammar

We design the K-J (or J-K) grammar which eliminates syntactical and semantical ambiguity of both languages for interpretation. This grammar corresponds to the communicative competence for the interpretation between Korean and Japanese. The K-J (J-K) grammar is motivated by grammar modification and the covering grammar.

ALGORITHM: irregularity categories removal or adjustment and semantic features insertion.

Input: a 5-tuple phrase structure grammar G = (N,Tk,Tj, P,S).

Output: an equivalent 5-tuple phrase structure grammar G' = (N',Tk'[sem_j],Tj',P',S').

Method: empirical and heuristic method.

Here N and N' are nonterminals, Tk, Tj, Tk' and Tj' are terminals, sem_j is semantic features, P and P' are production rules, and S and S' are the start symbols. The J-K grammar is designed analogously. In the framework of the generalized phrase structure grammar, the semantic features are accepted by a special phrase structure rule, that is a linking rule, which causes the relevant information about the phrase to be passed down the tree as a feature on the syntactic nodes. Therefore, interpretation procedure is constructed by a succinct algorithm founded on the K-J(J-K) grammar.

5. Fragments of Interpretation

In this section, we exhibit the fragments of our interpretation system: how phrase structure rules and semantic features interact in the interpretation procedure according to the K-J(J-K) grammar.

5.1 Homonymous construction

There are some kinds of construction types provided by syntax relations of each constituent. Among them, modification is a construction type related to Head and Attributes. Coordination implies that more than two subconstituents have syntactical coordination relation. Let us consider the following Japanese utterances:

1) 学校へ行っ［て］パンを食べる。 (modification)
`(Someone) goes to school, and eats bread.'

2) パンを食べ［て］学校へ行く。 (coordination)
`(Someone) eats bread and goes to school.'

The two utterances imply the semantic notions of modification and coordination, respectively, but have the same conjunction morpheme [te]. Semantically, they are represented in Korean by the outcome of interpretation as follows:

1) 학교에 가서 빵을 먹는다. (modification)

2) 빵을 먹고 학교에 간다. (coordination)

All such morpheme ambiguities induce not only lexical semantic ambiguity but sentential ambiguity. In order to interpret such ambiguous utterances, we employ semantic feature specification as the discipline of the semantic conjunction schemata. The following rules account immediately for the sentences in the example. Here we use the GPSG notations:

(1) modification schema

$$S \rightarrow H[sem_{\alpha 0}, \text{Conj 서}], H[sem_{\alpha 1}]$$

$$\text{where } \alpha i \in \{ (0,1), (0,0) \}$$

(2) coordination schema

$$S \rightarrow H[sem_{\alpha 0}, \text{Conj 고}], H[sem_{\alpha 1}]$$

$$\text{where } \alpha i \in \{ (1,0), (1,1) \}$$

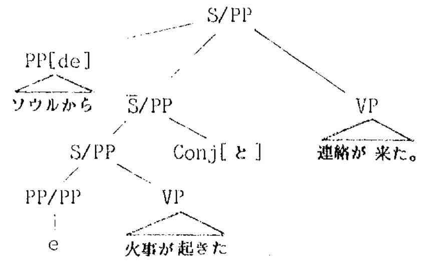
5.2 Missing construction

Korean and Japanese allow one of the constituents of a sentence not to be explicitly stated when it is understandable from the context. In the GPSG framework, this kind of difference can be expressed by a FOOT feature SLASH[3]. The SLASH feature indicates that something is missing in the structure dominated by the category specified. In this subsection, we exhibit a semantically ambiguous utterance across a homonymous construction and a missing construction. Consider the following Korean utterance. This utterance also has inherent syntactical and semantic ambiguity.

1) 서울에서 불이 났다고 연락이 왔다.

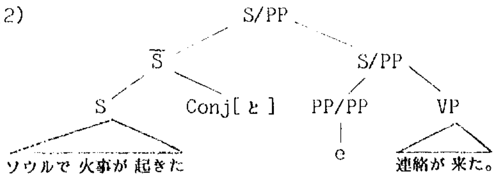
This utterance has two distinct syntactic trees:

(1)



`From Seoul came a report that there was a fire (in Seoul)'

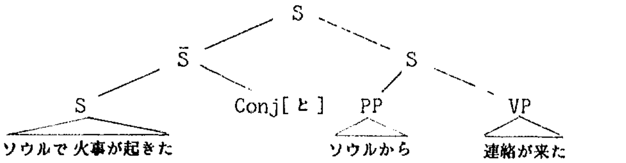
(2)



`(From Seoul) came a report that there was a fire in Seoul'

In the above example, homonymous construction does not arise in Japanese, but missing construction remains. We employ a parse tree (2) for semantic adjustment, and fill the gap of local environment with syntactically and semantically agreeable vocabulary; then such utterance of Korean and Japanese is interpretable without ambiguity. Consequently, the utterance of Korean 1) is interpreted as follows.

[[seoul- de [kazi-ga okitato] [seoul-kara [renraku ga kita]]].



'From Seoul came a report that there was a fire in Seoul'

6. K-J(J-k) system

In order to define a two-way interpretation system more formally, we formulate the internal interface(K-J system) for the interpretation. This interface corresponds to the transducer of interpretation. We can define the K-J(J-K) system as a 3-tuple grammar G=(w_j,k(or j),w_k), where w_k and w_j are Korean words and Japanese words, respectively, and k(j): $w_j \to w_k$ ($w_k \to w_j$) is a homomorphism. The K-J(J-K) system G defines the following sequence preserving the word order:

$$w_k^1 = k(w_j^1),\ \ w_k^1 w_k^2 = k(w_j^1)k(w_j^2),\cdots\cdots$$

It also defines the language

$$L(Gk) = \{k^i(w_j) | i>0\}.$$

As mentioned above, the K-J(J-K) system constitutes a simple device for interpretation. A language defined by the K-J(J-K) system corresponds to the target language. Inversely, the mapping j of w_k into w_j is such that the inverse homomorphism

$$j(w_k) = \{w_j | k(w_j) = w_k\}\ ,\ j = k^{-1}$$

exists. Thus, we define the two-way simultaneous interpretation system NARA by:

$$j(Lk) = k^{-1}(Lk) = \{w_j | k(w_j) \in Lk\}.$$

We can define our system NARA using the extended notion; the inverse homomorphism can be replaced by the direct operation of a finite substitution. Consider a grammar(e.g. Korean) Gk = (Nk, Tk, Pk, Sk) and let j be a finite substitution, defined on the vocabulary (Nk ∪ Tk)*, such that j(w) is a finite(possibly empty) set of word for each word w. We denote

$j(Nk) = Nj,\ j(Tk) = Tj,\ Pj \supset j(Pk),\ Sj \supset j(Sk).$ Then, the grammar(e.g. Japanese)

$$Gj = (Nj,\ Tj,\ Pj,\ Sj)$$

is an interpretation of Gk. If I(Gk), I(Gj) are the sets of all interpretation of Gk and Gj, respectively, then I(Gk) = I(Gj), and I is an invariant for Gk and Gj.

7. Complexity of System NARA

The complexity of the algorithm is usually measured by the growth rate of its time and space requirements, as a function of the size of its input (or the length of input string) to which the algorithm is applied. We adopt a finite state transducer as a computing model which governs the fundamental interpretation control. Since we do not count the time it takes to read the input, finite state languages have zero complexity. If reading the input is counted, then finite languages have time complexity of exactly n (the length of input string). Such languages are interpretable in exactly time n, and then called real-time languages. The interpretation which is accompanied by co-occurrence dependency cannot be done in general without relying on arbitrary look-ahead or rescanning of the output. However, the nature of on line interpretation is unchangeable. Consequently, our system NARA is interpreted in real-time.

8. Concluding Remarks

Our approach for constructing this system has both logical view and experimental view; the former is given by mathematical formalization, the latter by the correspondence of two languages. In the view of computational linguistics, we separated the mechanism of our two-way simultaneous interpretation system into the levels of abstract theory, algorithm, and implementation to carve out the results at each level in more independent fashion. In order to do so, we specified four important levels of description; the lowest level is morphology, the second level is segmented word, the third level is syntax and semantics, and the top level controls the computing model of each level. Hence, we could determine the range of correspondence between internal representations of both grammars, and the basic architecture of the machinery actually instantiates the algorithm. Consequently, our model produces the extra power by the proposed theory with multiple levels of representation and systematic mapping between the corresponding levels of two languages, because interpretation efficiency requires both functional and mathematical discussions. Nevertheless, the complete pragmatic interpretation still remains quite obscure. Finally, we confront the problem whether it is possible to construct a two-way simultaneous interpretation system between other two different language systems such as Japanese and English. We presuppose that the key point of problem-solving is in the study of universality and individuality between two given languages.

Acknowledgments

We are deeply grateful to Prof. H. YAMADA for his encouragement. We would like to thank Dr. A. ADACHI and Dr. K. HASHIDA, for many stimulating discussions and for detailed comments, and to Mr. Y. SHIRAI and Mr. I. FUJISHIRO for suggestions to improve the paper.

References

[1] N. CHOMSKY, Aspects of the Theory of Syntax, M.I.T. Press, Reading, 1963.

[2] H. S. CHUNG, Current Korean: Elementary Sentence Patterns and Structures, Komasholin, Reading, 1982(in Japanese).

[3] GAZDAR, KLEIN, PULLUM and SAG, Generalized Phrase Structure Grammar, Blackwell, Reading, 1985.

[4] H. HORZ, Eine Neue Invariante für Kontext-freie Sprachen, Theoretical Computer Science 11, 1980.

[5] H. R. LEWIS, C. H. PAPADIMITRIOU, ELEMENTS OF THE THEORY OF COMPUTATION, Prentice-Hall, Inc. Reading, 1981.

[6] A. NIHOLT, Context-Free Grammar: Cover, Normal Forms and Parsing, Springer, Reading, 1980.

[7] A. SALOMAA, Jewels of Formal Language Theory, Computer Science Press, Reading, 1981.