Linguistic Processing in Automatic Interpretation System between English-Korean Language Pair

K. S. Choi, S. M. Lee, Y. J. LEE*

Computer Science Department
Korea Advanced Institute of Science and Technology (KAIST)
373-1 Kusong-Dong Yusong-Ku Taejon, 305-701, KOREA
Automatic Interpretation section, ETRI*

Abstract This paper presents the linguistic processing for the Automatic Interpretation system between English/Korean language pair. We introduce two machine translation systems, each for English-to-Korean and Korean-to-English, describe the system configuration and several characteristics, and discuss the translation evaluation results.

1 Introduction

As a submodule of a bidirectional interpretation system for English and Korean language pair, we have developed a machine translation system for English-to-Korean and another translation system for Korean-to-English is under development. Both system works on Unix workstations. MATES is the acronym for MAchine Translation Environment System. It refers to the systems developed or under development by KAIST. Both of MATES/EK and MATES/KE can be used as a submodule of any other application system.

2 MATES for English-to-Korean

MATES/EK was developed from 1988 through 1992 as a prototype English-to-Korean translation system, and another trial for a PC version is under going. The project for MATES/EK resulted in not only the system itself but also in several useful tools such as Grammar Writing Language, GWL environment system and dictionary managing tools. Since 1992, several modifications to the system such as tuning the dictionaries, enlarging the grammar coverage, and strengthening the interface system have been done.

The main characteristics of MATES/EK are the following. First, we focused on the extendability. One of the major feature of a Language is the continuous change and advance. So a machine translation system need continuous modification to the linguistic knowledge of the language. MATES/EK provides powerful tools to write, modify, and extend the grammar and manipulate the dictionaries. These can make system development and management easy.

All the subsystems were implemented in C language so that transportation to the other computer system can be easy. It provides high portability.

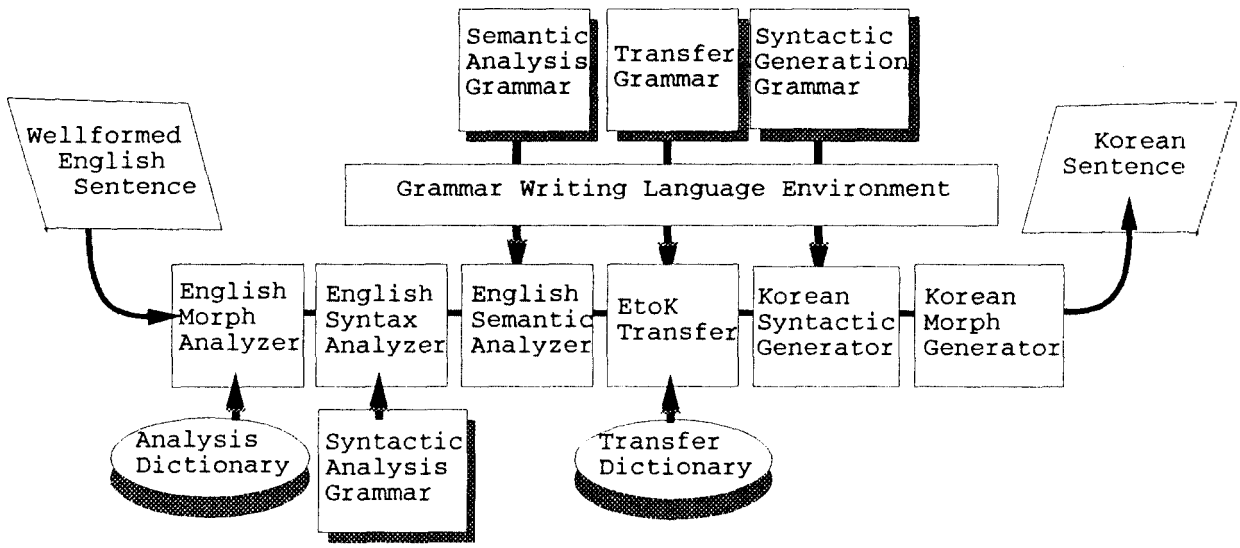

Figure 1: MATES/EK system configuration

Several kinds of ambiguities occur in translation process. The ambiguities scale down the system performance severely. In our system, we deal with the ambiguities in various levels. Morphological analyzer extracts proper complex word and performs categorial disambiguation based on trigram/rule table so that the search space of syntax analyzer can be reduced. In syntax analyzer, the grammar were divided by usage level which was based on the frequency of that rule. Analyzing a sentence using the levelized grammar generates the plausible(frequently used) trees first so that the number of parse trees can be also reduced. Also, Grammar Writing Language provides tree-scoring mechanism. These all can make the translation process more efficient and flexible.

2.1 System Configuration

The system configuration is depicted in figure 1. MATES/EK is based on transfer-based method consists of analysis phase, transfer phase, and generation phase. The intermediate representation was selected to effectively represent the meaning of source sentence and to simplify the transfer process. The intermediate representation, Lexical Semantic Structure(LSS) is a tree whose nodes are unorded 9 lexical-primitives and is less dependent on any language. There is only one relation between any two nodes which notifies what node is the head and what is the dependent node. These features of LSS makes itself appropriate for the intermediate representation between English and Korean whose syntactic structure are quite different.

The quality of translation is dependent on how plentiful, exact, and systemic the knowledge embedded in the dictionaries and the translation grammars is. In MATES/EK, there are two dictionary systems, one for analysis phase and the other for transfer and generation phases. Analysis dictionary has the information needed for morphological, syntactic, and semantic processing. Transfer dictionary is consists of morphological and syntactic knowledge for translation word selection and Korean sentence generation. Each dictionary in turn consists of domain-independent dictionary and domain-dependent dic-

tionary.

For development of several linguistic rules, we constructed a corpus consists of 3000 or so sentences most of which are from school texts and a lot of articles. The sentences plentiful of linguistic phenomena were chosen by linguistic experts. Morphological trigram/rules, syntax and semantic grammar and all other later-state grammars were inductively developed based on the corpus. For syntax analysis grammar, there are 460 Augmented Context Free rules. Semantic analysis grammar, transfer grammar, and syntax generation grammar are written in Grammar Writing Language. The grammars are composed of 296, 97, and 72 GWL-unit rules. Each GWL-unit rule corresponds to a tree-to-tree transformation. Both of ACFG rule and GWL-unit rule is augmented with several feature test conditions.

GWL have been used for semantic analysis, transfer, and generation grammar description. It's basic operation is tree-to-tree transformation. GWL consists of general grammar descriptor and transfer dictionary access rule descriptor and gives filtering and scoring mechanism for disambiguation. With a user-friendly notation, GWL makes grammar writing and maintenance an easy work.

2.2 Translation experiment

MATES/EK was experimented with about 1700 sentences selected in "IEEE Computer Magazine 91. September". The length of the 1700 sentences vary from 4 words through 24 words. We evaluated the translation result by 4 levels, each for excellent, good, poor, and fail. The criteria for the decision were degree of error-freeness and the acceptability test by human. Evaluation results are givien in the table 1. Sentences whose length is less than 14 words shows 95% success rate, sentences less than 18 words long shows about 90%, senetences less than 21 words long shows 80%, and the rests shows 75% success rate.

sentence length	excellent	good	poor	fail	total sentences
4 - 6	72	1	0	1	74
7 - 9	179	9	1	3	192
10 - 12	220	25	4	5	254
13 - 15	227	44	17	43	331
16 - 18	201	36	9	68	314
19 - 21	105	43	21	158	327
22 - 24	28	17	3	168	216

Table 1. Translation experiment result for English-to-Korean

"IEEE Computer Magazine" is consists of several articles of different authors. Accordingly the sentence styles of the articles were not uniform. And the sentences have a lot of complex linguistic phenomena such as enumeratives, inserted comments and so. Although such things scaled down the system performance, the result shown in the table 1 is encouraging.

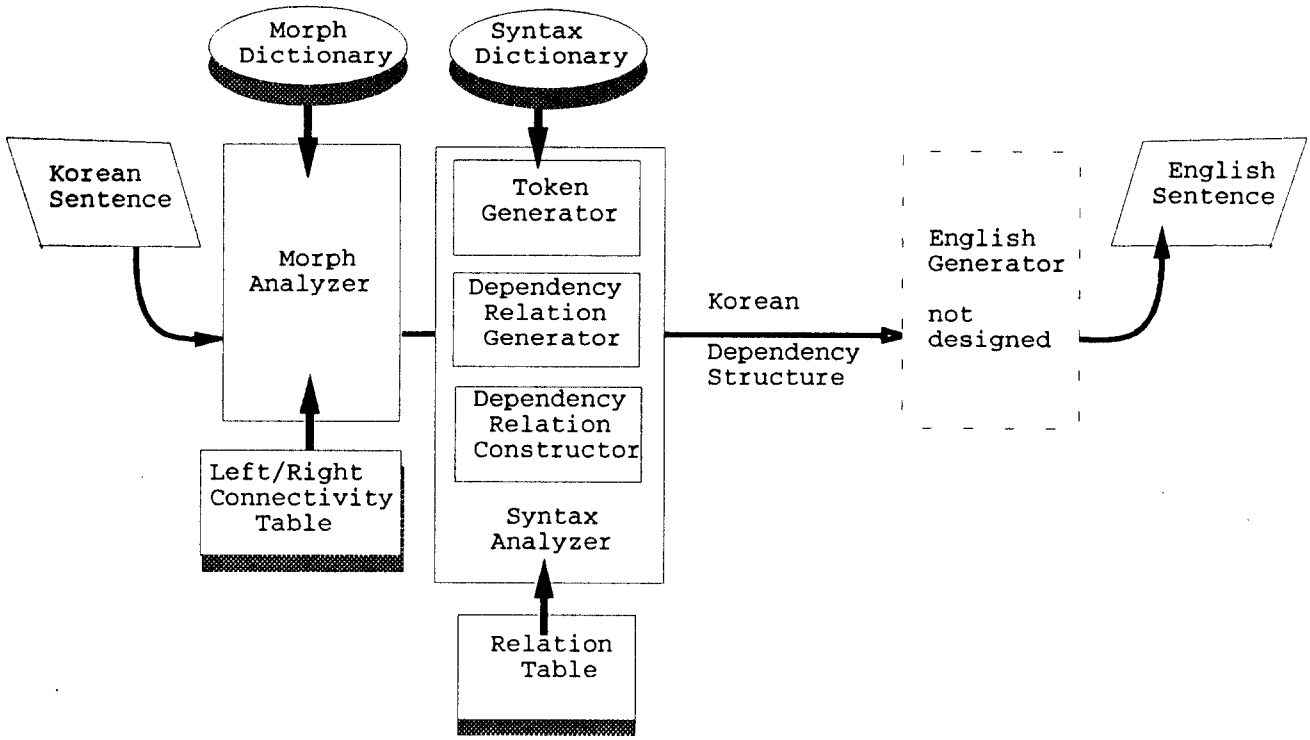

Figure 2: MATES/KE system configuration

Considering the MATES/EK as a submodule of an interpretation system, the translation performance seems to be better because dialogue sentences are usually less than 20 words. But the performance must also be evaluated on the ability of processing ill-formed input sentences. Currently MATES/EK assumes the input sentence to be well-formed. Our further research includes such robust translation mechanism and continuous enlarging/tuning the knowledge base.

3 MATES for Korean-to-English

MATES/KE is an under development system as a submodule of Automatic interpretation system for Korean-to-English. Currently, we are focusing on the development of Korean analysis technology. And for now, MATES/KE assumes the input sentence as a plain sentence style not a dialogue style which full of ellipsis and shorthand phenomena. In this section we introduce the general mechanism for Korean analysis.

3.1 Korean sentence Analysis

The system configuration for MATES/KE is depicted in figure 2. MATES/KE has two submodules of morphological analysis and syntax analysis. In turn, syntax analysis module consists of token generation phase, dependency relation construction phase, and dependency structure construction phase.

Korean sentences are consists of several Word-phrases which in turn consists of head word and its inflection codes, post position particles, and other affixes. So most of Korean morphological analyzers take an Word-phrases as its basic processing unit. The morphlogical analyzer of MATES/KE based on the longest-match analysis mechanism generates the morpheme list for each Word-phrases using the left/right connectivity information. The left/right connectivity table was constructed according to the connection possiblity of a category to the left and to the right of another category. Each morpheme in the generated list gets a fuzzy value which designates the appropriateness for each ambiguity.

A token is a unit on which a dependency relation can be exist. Tokens are basically a head category and other informations such as inflection and post position particles are encoded as feature information attached to the token. The process of dependency relation construction searches the possible dependency relation between two tokens and computes the degree of the relation strength using dependency information table and the distance of the two tokens in the sentence. Finally among the resulted dependency relations, A sequence of relation selection occurs: the strongest relation is selected if it results in no conflict with previously selected relations. In this way, we get the most plausible structures first. And for any ill-formed input sentences, we can get at least one structure which has lower relation strength between the tokens.

MATES/KE has two sub dictionaries for analysis phase, each for Korean morphological analysis and for Korean syntax analysis. Morphological analysis dictionary consists of about 60000 entries and gives information of left/right connectivity for each entry. For syntax analysis dictionary, there are about 5000 verbal entries which represent the case frame information.

3.2 Analysis experiment

We experimented the Korean analyzer with 150 sentences selected from a textbook of primary school. The analyzer always generates at least one plausible parse. So the evaluation was tested on the correctness of the parse. The results are shown in Table 2. There are two kinds of causes of any incorrectness. One is the cases that the morphological analyzer generates errorneous morpheme list(Type I error). And the other is the cases where the dependency relation construction occurs errorneous relations(Type II error). The average number of Word-phrases of test sentences were 5.9. The analysis success rate was about 75.3%. And for the 42 sentences of type I error, the success rate was 35.7%.

sentence length	# of sentences	# of correct analysis	success rate
less than 3 WPs	18	15	83.3%
4 – 5 WPs	54	44	81.4%
6 – 7 WPs	49	37	75.5%
8 – 9 WPs	19	11	57.9%
more than 10 WPs	10	6	60.6%
total	150	113	75.3%

Table 2. Analysis experiment for Korean

Future Research

In this paper, we presents two machine translation systems for English-to-Korean and for Korean-to-English as submodules of Automation Interpretation system. We introduce the MATES/EK and MATES/KE, describe main characteristics, and finally shows the translation experiment results.

Future work will be directed towards two main areas. One is the performance improvement through investigating ambiguity resolution mechanisms that cannot currently be handled in an elegant fashion. The other is related to the characteristics of a interpretation system. We will investigate the dialogue sentence style, robustness and etc. In addition, the design of English generation appropriate to MATES/KE will be investigated.

References

[1] Korea Advanced Institute of Science and Technology, *Research on English-to-Korean MT System (III): Development of Grammar Writing tools and English Analysis Grammars*, Korea Ministry Of Science and Technology, 1992

[2] Software Engineering Research Institute, *Research on Development of Linguistic Model and Skeleton System for English to Korean Machine Translation(3)*, Korea Ministry Of Science and Technology, 1992

[3] Korea Advanced Institute of Science and Technology, *Research on Development of Korean-to-English Machine Translation system for Automatic Interpretation*, Electronics and Telecommunications Research Institute, 1993

[4] Kwang-joon Seo, *A Korean Language Parser using Syntactic dependency relations between Word-phrases*, Korea Advanced Institute of Science and Technology M.S. thesis, 1992