AN ENGLISH-TO-KOREAN MACHINE TRANSLATOR : MATES/EK*

Key-Sun Choi, Seungmi Lee, Hiongun Kim, Deok-bong Kim,
Cheoljung Kweon, and Gilchang Kim

Center for Artificial Intelligence Research
Computer Science Department
Korea Advanced Institute of Science and Technology
Taejon, 305-701, Korea
{*kschoi, leesm, hgkim, dbkim, cjkwn, gckim*}*@csking.kaist.ac.kr*

Abstract

This note introduces an English-to-Korean Machine Translation System MATES/EK, which has been developed as a research prototype and is still under upgrading in KAIST (Korea Advanced Institute of Science and Technology). MATES/EK is a transfer-based system and it has several subsystems that can be used to support other MT-developments. They are grammar developing environment systems, dictionary developing tools, a set of augmented context free grammars for English syntactic analysis, and so on.

1. Introduction

An English-to-Korean machine translation system MATES/EK has been developed through a co-research done by KAIST and SERI (Systems Engineering Research Institute) from 1988 to 1992, and is still under evolution in KAIST. It has several tools supporting system development, such as the grammar writing language and its developing environment, a set of augmented context free grammar for English syntactic analysis, and dictionary editor. This system has been developed for UNIX workstation.

MATES/EK was originally developed using Common Lisp in order to test the possibility of English-to-Korean machine translation, and then it has been totally reconstructed using C language. Its main target domain is electric/electronic papers and so the dictionary and the grammars are specifically adjusted to the domain and one of sample sentences is IEEE computer magazine September 1991 issue to test and evaluate the system.

*This research is partly supported by Center for Artificial Intelligence Research (CAIR) (1992).

2. Overview of The System

MATES/EK is a typical transfer-based system, which does English sentence analysis, transforms the result (parse tree) into an intermediate representation, and then transforms it into a Korean syntactic structure to construct a Korean sentence. Figure 1 depicts the overall configuration of MATES/EK, which has following features:

- Morphological Analysis Using N-gram : We resolve the category ambiguities by combining the N-gram and the rules. (Kim, 1992)
- Augmented Context Free Grammars for English Syntactic Analysis : We developed a set of augmented context free grammar rules for general English syntactic analysis and the analyzer is implemented using Tomita LR parsing algorithm (Tomita, 1987).
- Lexical Semantic Structure (LSS) to represent the intermediate representation : The result of the syntactic structure is transformed into an intermediate representation LSS, which is a dependency structure that is relatively independent to specific languages. In LSS, the constituents in a sentence are combined only in head-dependent relation based on the lexical categories, and there are no order relation between the constituents. Hence LSS is desirable for trans-

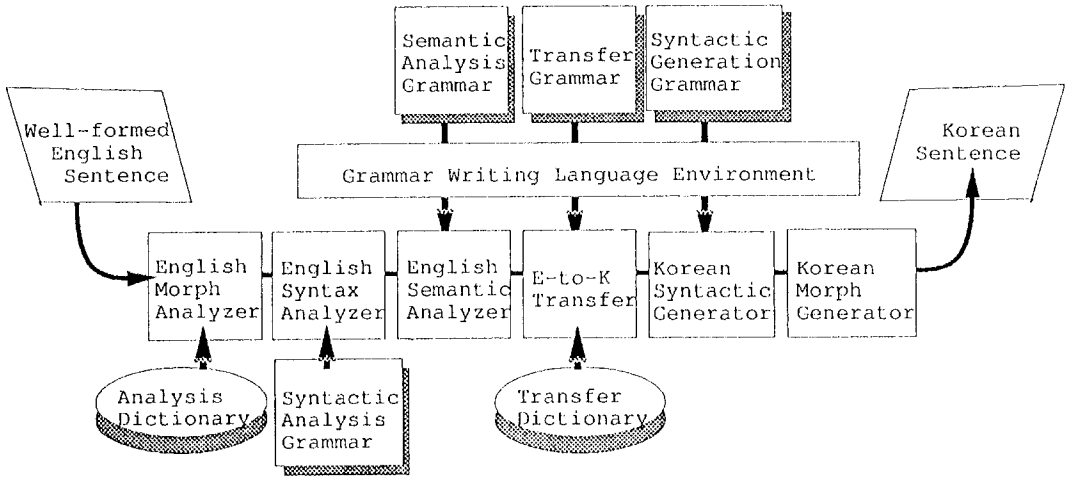

Figure 1: The System Configuration of MATES/EK

lation between English and Korean, two languages with fairly different syntactic structures (Kweon, et al., 1990, Kweon, 1992).

- Grammar Writing Language and Its Environment : MATES/EK runs a series of tree transformations on LSS structures from the English syntactic structure, in order to get a structure specific to Korean syntactic structure. To do this, a grammar writing language and the supporting system were developed for the tree transformations (Kweon, 1992).

  The whole tree transformations are done in a single grammar processing system in which a grammar writing language is defined and a set of tools, such as the parser, the interpreter and some debugging facilities for the language, are supported. In the grammar writing language, a rule describes a tree transformation by specifying the pattern of an input tree, test conditions, the transformation operations, and the resultant tree structures. Figure 2 is an example of a tree transformation rule written in grammar writing language and two trees before and after its application.

MATES/EK consists of a set of dictionaries, a set of grammar rules, and the processing modules. Translation is done through a series of processes; English morphological analysis, English syntactic analysis, English semantic analysis, English-Korean lexical transfer, English-to-Korean structural transformation, Korean syntactic structure generation, and Korean morphological generation. Brief introductions to each processing follows.

3. English Analysis

3.1. Morphological Analysis

It incorporates the method of categorial ambiguity resolution using N-gram with rule combinations, as well as the basic English word identification, such as word separation, processing of affixes and recognition of idiomatic phrases (Kim, et al., 1992).

3.2. English Syntactic Analysis

It uses the generalized Tomita LR parsing algorithm on augmented context free grammar. The grammar is inductively constructed from 3,000 carefully selected sentences that include various linguistic phenomena of English. Those sentences are mainly selected from the IEEE computer magazine September 1991. Other sources of the test sentences are HP issue test sentences, and Longman English Grammar Textbook. The constructed grammar for syntax analysis consists of about 500 rules.

As described above, LSS(the Lexical Semantic Structure) is description for the intermediate representation. The result of syntactic analysis is transformed into an LSS which is relatively more specific to English, and then is transformed into

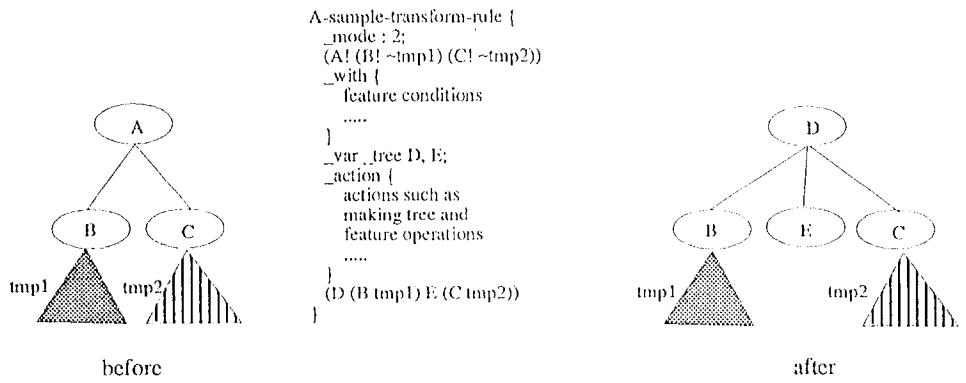

Figure 2: An example of grammar writing rule and the tree transformation – In the rule "(A! (B! tmp1) (C! tmp2))" describes that 'A' as a parent node of a pattern, '(B! tmp)' and '(C! tmp2)' as the first and second child, and each child may have zero or more children. The action part describes the necessary transformation operation.

an LSS specific to Korean.

3.3. English Semantic Analysis

We developed more than 300 tree transforming rules that are written in grammar writing language. These grammar rules lead the English syntactic structure into a dependency structure of English. This dependency structure is relatively similar to meaning structure but it is still specific to English, so we need more tree transformations to get a structure for Korean language.

4. English to Korean Transfer

In this step the system looks up the English-Korean bilingual dictionary. We manage the analysis dictionary separately from the transfer dictionary so that we may use the same analysis dictionary to the other language pair such as English to Japanese with the other transfer dictionary. There are more than 300 lexical specific selection rules developed to make the lexical selection better.

4.1. English-Korean Structural Transformation

Using another tree transformation grammar, the English specific dependency structure is transformed into a Korean language specific dependency structure after looking up the bilingual dictionary. The dependency structures are represented as head and dependents. Although the head in an English dependency structure is a English verb word, the head in corresponding Korean dependency structure is Korean verb or adjective word, those two words are often not mapped directly. Figure 3 is an example of transformation from an English syntactic structure into its English specific dependency structures LSS for a sentence "Pipelining increases performance by exploiting instruction-level parallelism."

5. Korean Generation

5.1. Korean Syntactic Generation

In this step the system transforms further the resultant structure into a list of Korean morphemes. Since the dependency structure specifies no word order, we have to find the word order of a sentence and necessary postpositions by help of rules and lexical information (Jang, et al., 1991). Note that Korean has, like Japanese, several alternative postpositions for conveying the same meaning.

5.2 Korean Morphological Generation

After the whole tree transformation, the resultant structure is a list of pairs of a morpheme and its category. The morphological generator is an automaton that does the synthesis and separation of morphemes according to the context and Korean morpheme combination rule. Since the Korean language has some complex morphological structure, the synthesis is a very complex process.

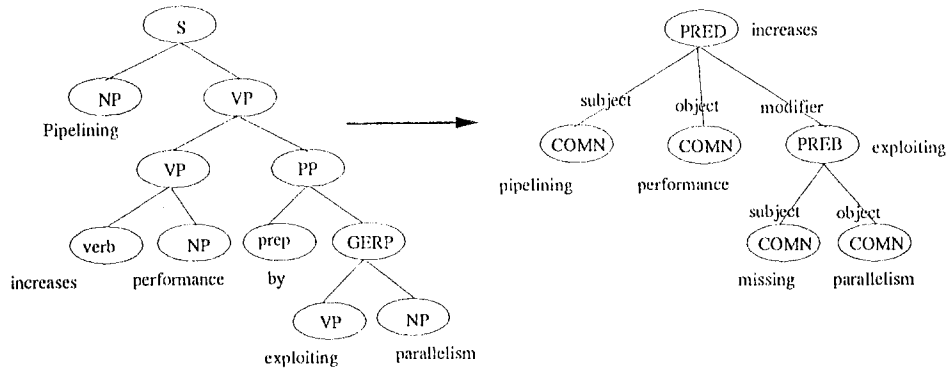

Figure 3: An example of English syntactic structure and the corresponding English dependency structure which is described in LSS, where; PRED (PREDicate) the head of a sentence, normally verbs or adjectives are selected, COMN (COMplement Noun) a node that leads a noun phrase, PREA (PREdicate Adjective) corresponds to a verb or an adjective in an adjective phrase, PREB (PREdicate adverB) corresponds to a verb or an adjective in an adverb phrase.

6. Problems for Evolution of the System

Since after the first completion of this project, we have been trying to find and solve the problems of this system. Following list is a brief list of those problems, and they are not seem to be easy to solve in the near future.

- Processing of non-continuous idiomatic expressions : In the dictionary entry specification we have a simple rule to represent the non-continuous idiomatic expressions, but it is not easy to detect those expressions from a sentence and represent the found expression in the internal structure for processing.

- Processing Korean Sentence Style : Korean language has various styles of sentences(difference between normal ones from the honorific or polite expressions), which are quite difficult to catch from the English sentences.

- Too many ambiguities in English syntactic analysis : Currently MATES/EK uses a set of ad hoc heuristics and lexical semantic markers coded in the dictionary in order to solve the ambiguity resolution, such as the PP attachment. This problem is related to the problem of selecting the right postposition of Korean.

- Robust processing for ill-formed sentences : Current MATES/EK assumes that the input sentence be a well formed English sentence. After practical test, we found the robustness for ill-formed sentences is highly required, because the papers from the IEEE computer magazine contains the non-sentential, non-text text such as braces, numeric expressions, formulas and so on.

- Selecting correct word correspondency between several alternatives : MATES/EK uses the semantic marker and a scoring of frequencies to select the word correspondency. The system still lacks a strong strategy for the word selection.

7. Test and Evaluation

Evaluation of an MT system emerges as a critical issue these days, but we have not yet found a strong and objective way of evaluation. After the first completion of the project we tried though, to make an evaluation of the system.

In order to make the evaluation as objective as possible we prepared three factors. First, the referees of the evaluation should be those who are not the developers of the system, and they should take a training to make objective decisions. We selected randomly five master degree students as the referees. Second, the referees are given a decision criteria of four levels: best, good, poor, and

fail. A sentence is 'best' translated if the resultant Korean sentence is very natural and requires no additional postediting. A sentence is 'good' translated if the result sentence is structurally correct but it has some minor lexical selection errors. A sentence is translated 'poor' if there is structural error as well as lexical errors. By 'fail', we mean when the system produces very ill-formed translation or fails to produce any result at all. We took the first three levels to be 'success,' because even a sentence is translated in 'poor' degree, it is still understandable. (Even if a translation is scored to be 'fail', it could sometimes be understandable from the view point of 'mechanical translation.') Third, the test sentences should be those sentences which were never used during the development time.

This system was tested on 1,708 sentences, whose length were less than 26 words selected from 2500 sentences in the IEEE computer magazine September 1991 issue. It showed about 95 percent of success rate for sentences shorter than 15 words, about 90 percent for 18 words, 80 percent for 21 words, and 75 percent for 26 words. This is a quite encouraging result since the IEEE computer magazine contains wide range of texts of various styles.

8. Conclusion and Further Study

Development of MATES/EK gave a strong motivation of attacking practically important problems, such as dictionary management, scaling up the grammar rules to the real texts, controlling the consistency of a large system.

The system MATES/EK is still under growing, trying to overcome the problems listed above, scaling up the dictionaries and the grammar rules, and doing downsizing to the PC environment.

References

[1] Choi, K.S., (1988). Developing Linguistic Model and Skeleton System for Machine Translation. *KAIST TR, M20160.*

[2] Choi, K.S., (1989). Research on English-to-Korean Transfer and Development of Machine Dictionary. *KAIST TR M03330.*

[3] Jang, M.G., et al., (1991). Korean Generation in MATES/EK. *Proceedings of Natural Language Processing Pacific Rim Symposium (NLPRS '91), Singapore.*

[4] Kim, D.B., Chang, D.S., and Choi, K.S., (1992). English Morphological Analyzer in English-to-Korean Machine Translation System. *PRICAI'92, Seoul.*

[5] Kweon, C.J., Choi, K.S., and Kim, G.C., (1990). Grammar Writing Language (GWL) in MATES-EK. *Proceedings of PRICAI 1990, Nagoya, Japan, November 14th 1990.*

[6] Kweon, C.J., (1992). Grammar Writing Language : CANNA-2. *KAIST TR, M40071.*

[7] Tomita, M., (1987). An efficient augmented-context free parsing algorithm. *Computational Linguistics. 13 (1-2), 1-6 1987.*