Hybrid Approaches to Improvement of Translation Quality in Web-based English-Korean Machine Translation

Sung-Kwon Choi, Han-Min Jung,
Chul-Min Sim, Taewan Kim, Dong-In Park
MT Lab. SERI
1 Eoun-dong, Yuseong-gu,
Taejon, 305-333, Korea
{skchoi, jhm, cmsim, twkim, dipark}@seri.re.kr

Jun-Sik Park, Key-Sun Choi
Dept. of Computer Science, KAIST
373-1 Kusong-dong, Yuseong-gu,
Taejon, 305-701, Korea
jspark@world.kaist.ac.kr
kschoi@cs.kaist.ac.kr

Abstract

The previous English-Korean MT system that was the transfer-based MT system and applied to only written text enumerated a following brief list of the problems that had not seemed to be easy to solve in the near future : 1) processing of non-continuous idiomatic expressions 2) reduction of too many ambiguities in English syntactic analysis 3) robust processing for failed or ill-formed sentences 4) selecting correct word correspondence between several alternatives 5) generation of Korean sentence style. The problems can be considered as factors that have influence on the translation quality of machine translation system. This paper describes the symbolic and statistical hybrid approaches to solutions of problems of the previous English-to-Korean machine translation system in terms of the improvement of translation quality. The solutions are now successfully applied to the web-based English-Korean machine translation system "FromTo/EK" which has been developed from 1997.

Introduction

The transfer-based English-to-Korean machine translation system "MATES/EK " that has been developed from 1988 to 1992 in KAIST(Korean Advanced Institute of Science and Technology) and SERI(Systems Engineering Research Institute) enumerated following list that doesn't seem to be easy to solve in the near future in terms of the problems for evolution of the system (Choi *et. al.*, 1994) :

- processing of non-continuous idiomatic expressions
- generation of Korean sentence style
- reduction or ranking of too many ambiguities in English syntactic analysis
- robust processing for failed or ill-formed sentences
- selecting correct word correspondency between several alternatives

The problems result in dropping a translation assessment such as fidelity, intelligibility, and style (Hutchins and Somers, 1992). They can be the problems with which MATES/EK as well as other MT systems have faced.

This paper describes the symbolic and statistical hybird approaches to solve the problems and to improve the translation quality of web-based English-to-Korean machine translation.

1 System Overview

English-to-Korean machine translation system "FromTo/EK" has been developed from 1997, solving the problems of its predecessor "MATES/EK" and expanding its coverage to WWW. FromTo/EK has basically the same formalism as MATES/EK that does English sentence analysis, transforms the result (parse

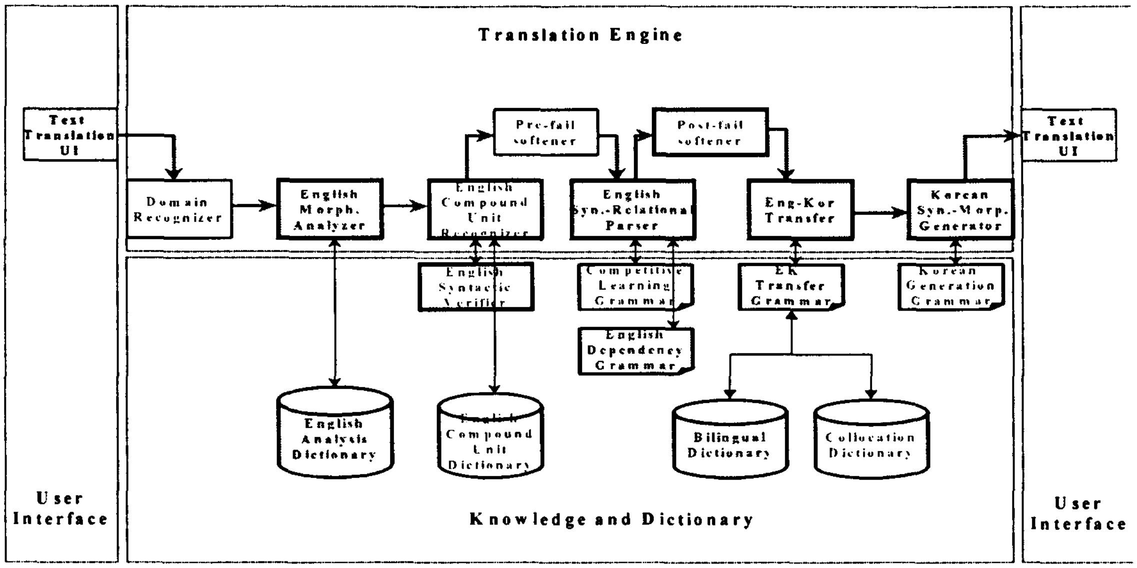

Figure 1: The System Configuration of FromTo/EK

tree) into an intermediate representation, and then transforms it into a Korean syntactic structure to construct a Korean sentence. Figure 1 shows the overall configuration of FromTo/EK. FromTo/EK consists of user interface for English and Korean, translation engine, and knowledge and dictionaries. The black boxes in the Figure 1 mean the modules that have existed in MATES/EK, while the white ones are the new modules that have been developed to improve the translation quality. Next chapters describe the new modules in detail.

2 Domain Recognizer and Korean sentence style

In order to identify the domain of text and connect it to English terminology lexicon and Korean sentence style in Korean generation, we have developed a domain recognizer.

We adapted a semi-automated decision tree induction using C4.5 (Quinlan, 1993) among diverse approaches to text categorization such as decision tree induction (Lewis *et. al.*, 1994) and neural networks (Ng *et. al.*, 1997), because a semi-automated approach showed perhaps the best performance in domain identification according to (Ng *et. al.*, 1997). Twenty-five domains were manually chosen from the categories of awarded Web sites. We collected 0.4 million Web pages by using Web search robot and counted the frequency of words to extract features for domain recognition. The words that appeared more than 200 times were used as features. Besides we added some manually chosen words to features because the features extracted automatically were not able to show the high accuracy.

Given an input text, our domain recognizer assigns one or more domains to an input text. The domains can raise the translation quality by activating the corresponding domain-specific terminology and selecting the correct Korean sentence style. For example, given a "driver", it may be screw driver, taxi driver or device driver program. After domain recognizer determines each domain of input text, "driver" can be translated into its appropriate Korean equivalent. The domain selected by the domain recognizer is able to have a contribution to generate a better Korean sentence style because Korean sentence style can be represented in various ways by the verbal endings relevant to the domain. For example, the formal domains such as technology and law etc. make use of the plain verbal ending like 'ta' because they have carateristics of formality, while the informal domains such as weather, food and fashion etc. are related to the polite verbal ending 'supnita' because they have carateristics of politeness.

3 Compound Unit Recognition

One of the problems of rule-based translation has been the idiomatic expression which has been dealt mainly with syntactic grammar rules (Katoh and Aizawa, 1995) "Mary keeps up with her brilliant classmates." and "I prevent *him* from *go*ing there." are simple examples of uninterupted and interupted idiomatic expressions expectively.

In order to solve idiomatic expressions as well as collocations and frozen compound nouns, we have developed the compound unit(CU) recognizer (Jung *et. al.,* 1997). It is a plug-in model locating between morphological and syntactic analyzer. Figure 2 shows the structure of CU recognizer.

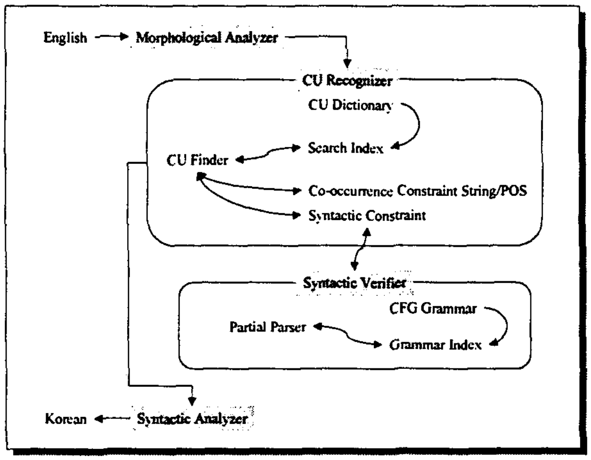

Figure 2 : System structure of CU recognizer

The recognizer searches all possible CUs in the input sentence using co-occurrence constraint string/POS and syntactic constraint and makes the CU index. Syntactic verifier checks the syntactic verification of variable constituents in CU. For syntactic verifier we use a partial parsing mechanism. Partial parser operates on cyclic trie and simple CFG rules for the fast syntactic constraint check. The experimental result showed our syntactic verification increased the precision of CU recognition to 99.69%.

4 Competitive Learning Grammar

For the parse tree ranking of too many ambiguities in English syntactic analysis, we use the mechanism to insert the competitive probabilistics into the rules. To decide the correct parse tree ranking, we compare two partial parse trees on the same node level with competitive relation and add α (currently, 0.01) to the better one, but subtract α from the worse one on the base of the intuition of linguists. This results now in raising the better parse tree higher in the ranking list of the parse trees than the worse one.

5 Robust Translation

In order to deal with long sentences, parsing-failed or ill-formed sentences, we activate the robust translation. It consists of two steps: first, long sentence segmentation and then fail softening.

5.1 Long Sentence Segmentation

The grammar rules have generally a weak point to cover long sentences. If there are no grammar rules to process a long sentence, the whole parse tree of a sentence can not be produced. Long sentence segmentation produces simple fragements from long sentences before parsing fails.

We use the POS sequence of input sentence as a clue of the segmentation. If the length of input sentence exceeds pre-defined threshold, currently 21 for segmentation level I and 25 for level II, the sentence is divided into two or more parts. Each POS trigram is separately applied to the level I or II. After segmenting, each part of input sentence is analyzed and translated. The following example shows an extremely long sentence (45 words) and its long sentence segmentation result.

[Input sentence]
"Were we to assemble a Valkyrie to challenge IBM, we could play Deep Blue in as many games as IBM wanted us to in a single match, in fact, we could even play multiple games at the same time. Now - - wouldn't that be interesting?"

[Long Sentence Segmentation]
"Were we to assemble a Valkyrie to challenge IBM, / (noun PUNCT pron) we could play Deep Blue in as many games as IBM wanted us to in a single match, / (noun PUNCT adv) in fact, / (noun PUNCT pron) we could even play multiple games at the same time, / (adv PUNCT adv) Now - - / (PUNCT PUNCT aux) wouldn't that be interesting?"

5.2 Fail Softening

For robust translation we have a module 'fail softening' that processes the failed parse trees in case of parsing failure. Fail softening finds set of edges that covers a whole input sentence and makes a parse tree using a virtual sentence tag. We use left-to-right and right-to-left scanning with "longer-edge-first" policy. In case that there is no a set of edges for input sentence in a scanning, the other scanning is preferred. If both make a set of edges respectively, "smaller-set-first" policy is applied to select a preferred set, that is, the number of edges in one set should be smaller than that of the other (e.g. if n(LR)=6 and n(RL)=5, then n(RL) is selected as the first ranked parse tree, where n(LR) is the number of left-to-right scanned edges, and n(RL) is the number of right-to-left scanned edges). We use a virtual sentence tag to connect the selected set of edges. One of our future works is to have a mechanism to give a weight into each edge by syntactic preference.

6 Large Collocation Dictionary

We select a correct word equivalent by using lexical semantic marker as information constraint and large collocation dictionary in the transfer phase.

The lexical semantic marker is applied to the terminal node for the relational representation, while the collocation information is applied to the non-terminal node.

The large collocation dictionary has been collected from two resources; EDR dictionary and Web documents.

7 Test and Evaluation

A semi-automated decision tree of our domain recognizer uses as a feature twenty to sixty keywords which are representative words extracted from twenty-five domains. To raise the accuracy of the domain identifier, manually chosen words has been also added as features.

For learning of the domain identifier, each thousand sentence from twenty-five domains is used as training sets. We tested 250 sentences that are the summation of each ten sentences extracted from twenty-five domains. These test sentences were not part of training sets. The domain identifier outputs two top domains as its result. The accuracy of first top domain shows 45% for 113 sentences. When second top domains are applied, the accuracy rises up to 75%.

In FromTo/EK, the analysis dictionary consists of about 70,000 English words, 15,000 English compound units, 80,000 English-Korean bilingual words, and 50,000 bilingual collocations. The domain dictionary has 5,000 words for computer science that were extracted from IEEE reports.

In order to make the evaluation as objective as possible we compared FromTo/EK with MATES/EK on 1,708 sentences in the IEEE computer magazine September 1991 issue, which MATES/EK had tested in 1994 and whose length had been less than 26 words. Table 1 shows the evaluation criteria.

Table 1: The evaluation criteria

Degree	Meaning
4 (Perfect)	The meaning of the sentence is perfectly clear.
3 (Good)	The meaning of the sentence is almost clear.
2 (OK)	The meaning of the sentence can be understood after several readings.
1 (Poor)	The meaning of the sentence can be guessed only after a lot of readings.
0 (Fail)	The meaning of the sentence cannot be guessed at all.

With the evaluation criteria three master degree students whom we randomly selected compared and evaluated the translation results of 1,708 sentences of MATES/EK and those of FromTo/EK. We have considered the degrees 4, 3, and 2 in the table 1 as successful translation results. Figure 3 shows the evaluation result.

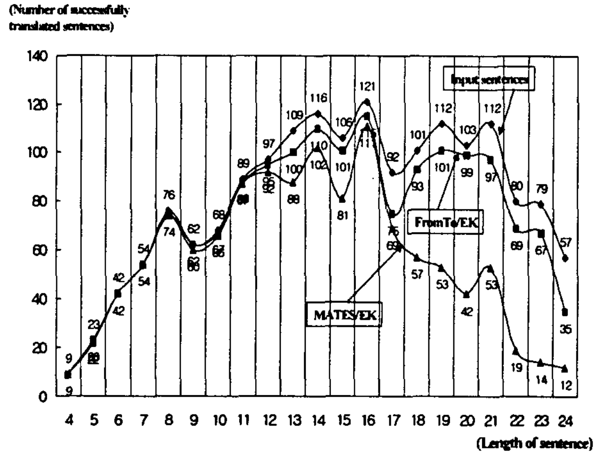

Figure 3 : The evaluation of 1,708 sentences

Figure 3 shows a translation quality of both FromTo/EK and MATES/EK according to the length of a sentence. More than 84% of sentences that FromTo/EK has translated is understood by human being.

8 Conclusion

In this paper we described the hybrid approaches to resolution of various problems that MATES/EK as the predecessor of FromTo/Ek had to overcome. The approaches result in improving the translation quality of web-based documents.

FromTo/EK is still under growing, aiming at the better Web-based machine translation, and scaling up the dictionaries and the grammatical coverage to get the better translation quality.

References

Choi K.S., Lee S.M., Kim H.G., and Kim D.B. (1994) *An English-to-Korean Machine Translator: MATES/EK.* COLING94, pp. 129-133.

Hutchins W.J. and Somers H.L. (1992) *An Introduction to Machine Translation.* Academic Press.

Jung H.M., Yuh S.H., Kim T.W., and Park D.I. (1997) *Compound Unit Recognition for Efficient English-Korean Translation.* Proceedings of ACH-ALLC.

Katoh N. and Aizawa T. (1995) *Machine Translation of Sentences with Fixed Expression.* Proceedings of the 4th Applied Natural Language Processing.

Lewis D.D. and Ringuette M. (1994) *A comparison of two learning algorithms for text categorization.* Symposium on Document Analysis and Information Retrieval, pp.81-93.

Ng H., Goh W., and Low K. (1997) *Feature Selection, Perceptron Learning, and a Usability Case Study for Text Categorizatio.* Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Quinlan J. (1993) *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers.