English-to-Korean Web Translator: "FromTo/Web-EK"

Sung-Kwon Choi, Taewan Kim, Sang-Hwa Yuh, Han-Min Jung, Chul-Min Sim, Sang-Kyu Park
Knowledge Processing Research Team
Electronics and Telecommunications Research Institute
Korea

Abstract.

The previous English-Korean MT system that have been developed in Korea have dealt with only written text as translation object. Most of them enumerated a following 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 POS or structural ambiguities 3) robust processing for long sentence and parsing failure 4) selecting correct word correspondence between several alternatives. The problems can be considered as important factors that have influence on the translation quality of machine translation system. This paper describes not only the solutions of problems of the previous English-to-Korean machine translation systems but also the HTML tags management between two structurally different languages, English and Korean. Through the solutions we translate successfully English web documents into Korean one in the English-to-Korean web translator "FromTo/Web-EK" which has been developed from 1997.

1 Introduction

The huge growth of the Internet allows human to get the unrestricted useful information from Internet. But the problem is always the language barrier, and becomes worse between structurally different language group, as we know well.

In order to solve the language barrier the machine translation systems from either English or Japanese to Korean or reverse have been developed actively in Korea since 1987. The previous English-to-Korean MT systems that have been developed in Korea have dealt with only written text as translation object. Most of them enumerated a following list of the problems that had not seemed to be easy to solve in the near future in terms of the problems for evolution of the system (Choi *et al,* 1994):

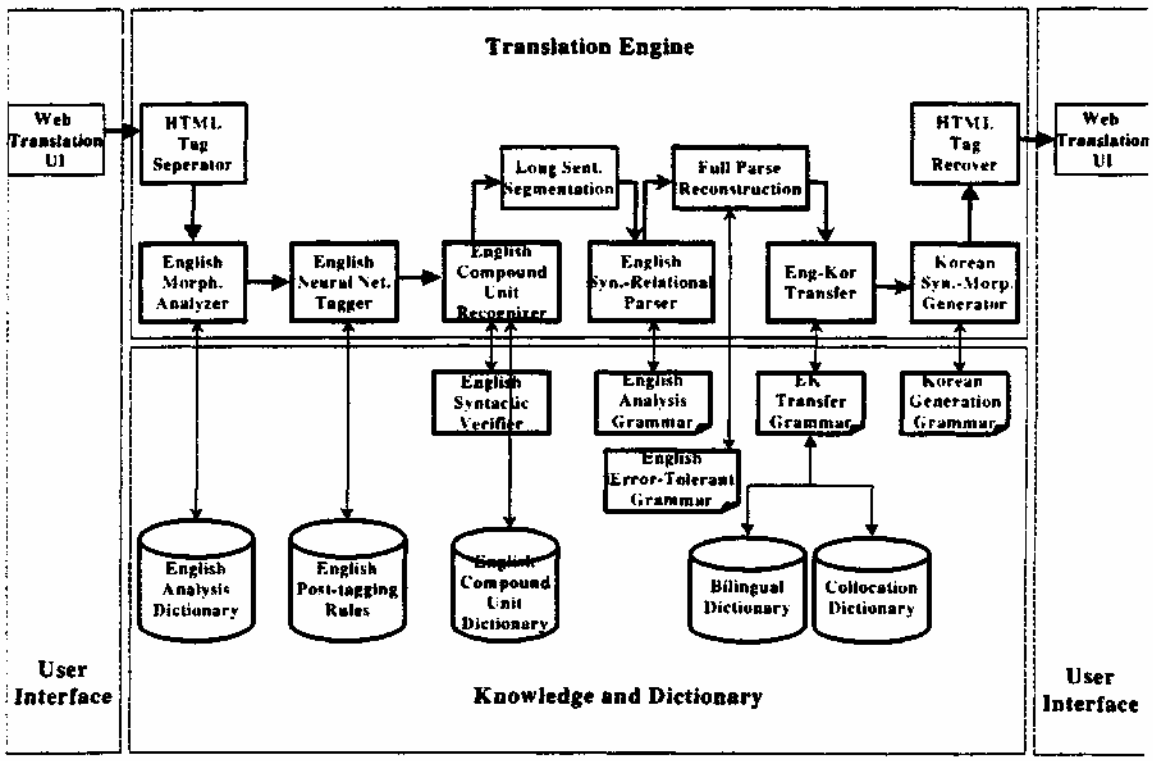
- processing of non-continuous idiomatic expressions
- reduction of too many ambiguities of POS or structural ambiguities
- robust processing for long sentences and failed or ill-formed sentences
- selecting correct word correspondency between several alternatives

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 previous English-to-Korean MT systems as well as other MT systems also have faced.

This paper describes not only the solutions of problems of previous English-to-Korean machine translation systems but also the methods transferring the HTML tags for web-based machine translation on Internet between two structurally different languages such as English and Korean. This paper is written on the basis of the English-to-Korean web translator "FromTo/Web-EK" which has been developed from 1997.

2 System Overview

English-to-Korean web translator FromTo/Web-EK has been developed from 1997 to 1998, solving the problems of machine translation systems for written text and expanding its coverage to WWW. FromTo/Web-EK belongs to the rule-based methodology for machine translation and has tree transduction formalism that does English sentence analysis, transforms the result (parse tree) into an intermediate representation, and then transforms it



[Figure 1] The System Configuration of FromTo/Web-EK

into a Korean syntactic structure to construct a Korean sentence. Figure 1 shows the overall configuration of FromTo/Web-EK. FromTo/Web-EK consists of three parts : user interface for English and Korean, translation engine, and knowledge and dictionaries. Next chapters describe modules in detail.

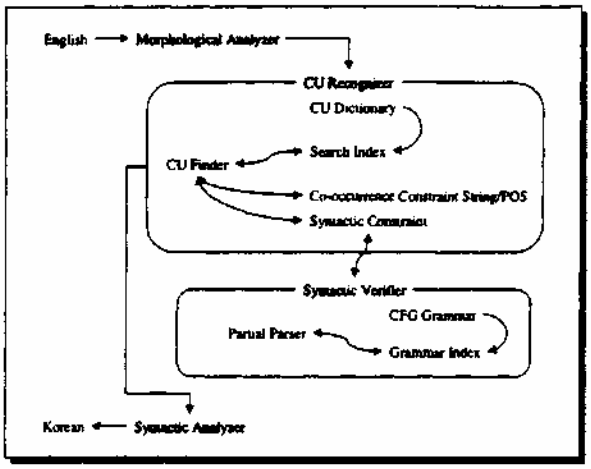
3 Neural Network Tagger with Post-tagging rules

So far, the following various models for POS tagging have been proposed : transformation-based tagging (Brill, 1992), neural network model (Schmid, 1994), Hidden Markov Model (HMM) (Kupiec, 1992), and recently a Maximum Entropy Model (Ratnaparkhi, 1996). The accuracy rate of these models varies from 95% to 96.6%. We propose a hybrid N-best English tagger (Yuh et al, 1999). It attaches post tagging rules to a neural network tagger. The post tagging rules solve the POS ambiguity of the tagger's output. We evaluated our tagger with both HMM tagger and genuine neural network tagger. The experimental results of the tagging accuracy for two kinds of 2,000 sample sentences show 97.5%, which shows that our neural network tagger with post tagging rules outperforms both HMM tagger and genuine neural network tagger by 1.4% and 1.7%, respectively. It means that our tagger realizes in advance a reduction of ambiguities in English syntactic analysis. Now we have 117 post tagging rules in relation of converting morphological part-of-speech.

4 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 going there.” are simple examples of uninterrupted and interrupted idiomatic expressions respectively.

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



[Figure 2] System structure of Compound Unit Recognizer

The recognizer searches all possible CUs in the tagger's output 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%.

5 Robust Translation with Long Sentence Segmentation and Full Parse Reconstruction

In order to deal with long sentences and parsing failure, we activate the robust translation. It consists of two steps: first, long sentence segmentation and then full parse reconstruction.

5.1 Long Sentence Segmentation

The grammar rules as translation knowledge have generally a weak point to cover long sentences because they can cause chart over flow due to structural ambiguities. Long sentence segmentation may prevent in advance such structural ambiguities so that it produces simple fragments from long sentences before parsing fails.

We use the POS sequence of tagger's output and its feature as a clue of the segmentation. If the length of input sentence exceeds pre-defined threshold considered by word order of segmented fragments according to the length of words in a sentence, currently 15 words for segmentation level I, 20 words for segmentation level II and 25 words for level III, a sentence can be divided into two or more pans. Each POS trigram is separately applied to the level I, II or III. Now we have 157 rules for long sentence segmentation. 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]

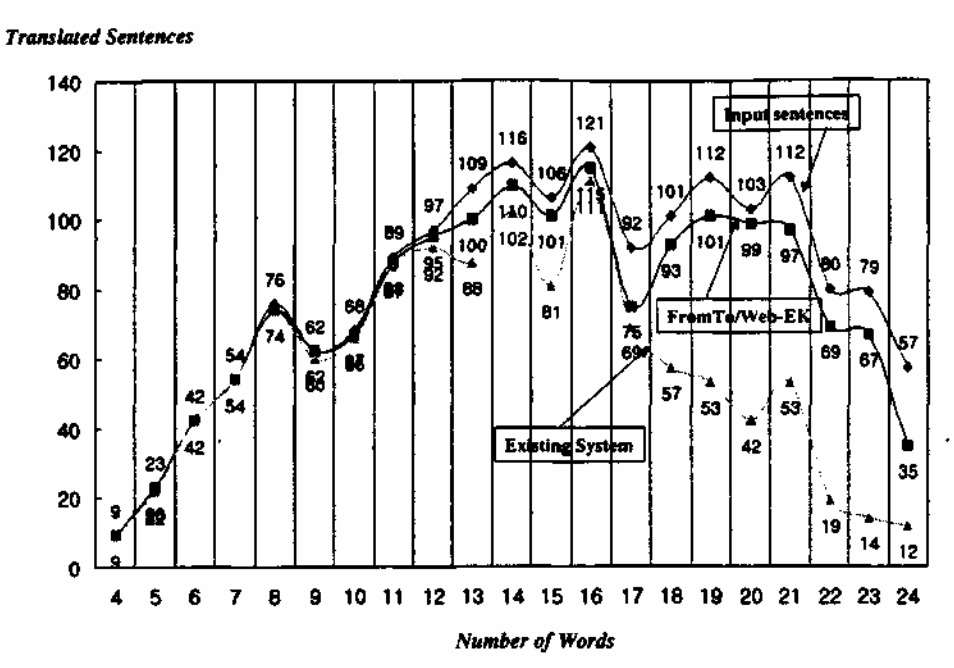
"Were we to assemble a Valkyrie to challenge IBM, / (noun PUNCT pron) we could play Deep Blue in as many games as IBM wanted us to in a single match, / (noun PUNCT adv) in fact, / (noun PUNCT pron) we could even play multiple games at the same time, / (adv PUNCT adv) Now - - / (PUNCT PUNCT aux) wouldn't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

5.2 Full Parse Reconstruction

For robust translation we have a module 'full parse reconstruction' that reconstructs the whole parse tree with partially successful parse trees in case of parsing failure by using error-tolerant grammar with 81 rules. Full parse reconstruction 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 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.

6 Collocation Dictionary and Lexical Rules

We are connecting collocation dictionary with lexical rules to select a correct word equivalent. The entries of collocation dictionary are being collected as large corpus from two resources : EDR dictionary and Web documents. The lexical rules have been made to support coverage of collocation dictionary



[Figure 3] The evaluation of 1,708 sentences.

and are lexical semantics-oriented with 43 lexical semantic codes. They are related to semantic marker between a governing non-terminal node and its dependents. Now we have 169 lexical rules.

7 HTML tags Management

The HTML tags themselves are not the translation objects, but they should be maintained in the appropriate position after the translation. Otherwise, structurally different languages such as English and Korean may have very more different layout than expected due to divergency of word order after translation. Therefore, the HTML tag management should be devised precisely to maintain the document layout and link information after the translation. In our system it consists of two phases: tag separation and tag recovery.

7.1 Tag Separation

In the tag separation phase, we use the layout information and the punctuation marks of HTML documents. Our tag separation strategy is as follows;

a) Pairs of start tag and end tag should be reserved (e.g. <A>, </A>, <TITLE>, </TITLE>)
b) If there are several sentences within a tag pair, they are separated by the punctuation marks.
c) Each item within a cell of table is regarded as a sentence.
d) Each item of a list is regarded as a sentence.
e) A compound noun and a phrase that are followed by blank line are regarded as a title.

The HTML tags are expected and stored in an external file with the sentence number (SN), the word number (WI), word (WD), start-tag (ST), end tag (ET), and flag information (FI).

7.2 Tag Recovery

If the target words are matched by 1 to 1, there are no problems for tag recovery. However, 1 to n, n to 1, or n to m transfer needs tag expansion and integration. During the translation, translator handles not full tags, but just word sequence information or token Ids. The tag manger maintains the whole tags. After the document is fully translated, the tag manager recovers HTML tags according to the following tag recovery strategy :

a) 1 to n: start tags and end tags of source language word are simply copied to target Korean word.
b) n to 1 : tag manager analyses start tags and end tags of source words. It determines tags which must be preserved in a clue of the sequencing information (e.g. <A>, </A>, <TITLE>,

</TITLE>). During this phase, internal tags including sizing and colouring tags are ignored. c) n to m : usually, in the case of idioms or compound units n to m transfer appears. Like the n to 1 case, the tag manager analyses tags of source words. Then it decides start tags and end tags of target words. During this phase, some tags of colouring or sizing can be excluded.

8 Experiment and Evaluation

In "FromTo/Web-EK" the dictionary consists of about 200,000 English full-form words with weight for neural network tagging, about 70,000 English lexeme for English analysis, about 22,000 English-Korean compound units for English frozen expression recognition, 80,000 English-Korean bilingual lexeme for transfer, and 50,000 bilingual collocations. Now different terminology is being constructed according to the domain.

In order to make the evaluation as objective as possible we have used the following evaluation criteria which is decided by human translator.

[Table 1] The evaluation criteria

Criterium	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.

On the basis of the evaluation criteria in Table 1, three students with English master degree whom we randomly selected compared and evaluated the translation results of an existing English-to-Korean machine translation system and those of "FromTo/Web-EK". The evaluation data were 1,708 sentences in the IEEE computer magazine September 1991 issue, which an existing English-to-Korean machine translation system had tested in 1994 and whose length had been less than 26 words. Figure 3 shows the evaluation result. In Figure 3 the upper line indicates total number of input sentences sorted by the number of words of a sentence. The middle line shows the number of sentences translated by our web translator and the low line is the translated result of existing English-to-Korean machine translation system.

We have considered the degrees 4, 3, and 2 in the table 1 as successful translation results. According to Figure 3, we know that more than 84% of sentences that our web translator "FromTo/Web-EK" has translated are also understood by human translator.

9 Conclusion

In this paper we described the English-to-Korean web translator "FromTo/Web-EK" that has solved various problems that existing English-to-Korean machine translation systems as well as most of rule-based machine translation systems had to overcome. The approaches resulted in improving the translation quality of web documents. FromTo/Web-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.

Figure 4 shows an example translated by English-to-Korean Web Translator "FromTo/Web-EK".

References

Brill, E. (1992). A Simple Rule-based Part of Speech Tagger, Proceedings of the DARPA Speech and Natural Language Workshop.

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.

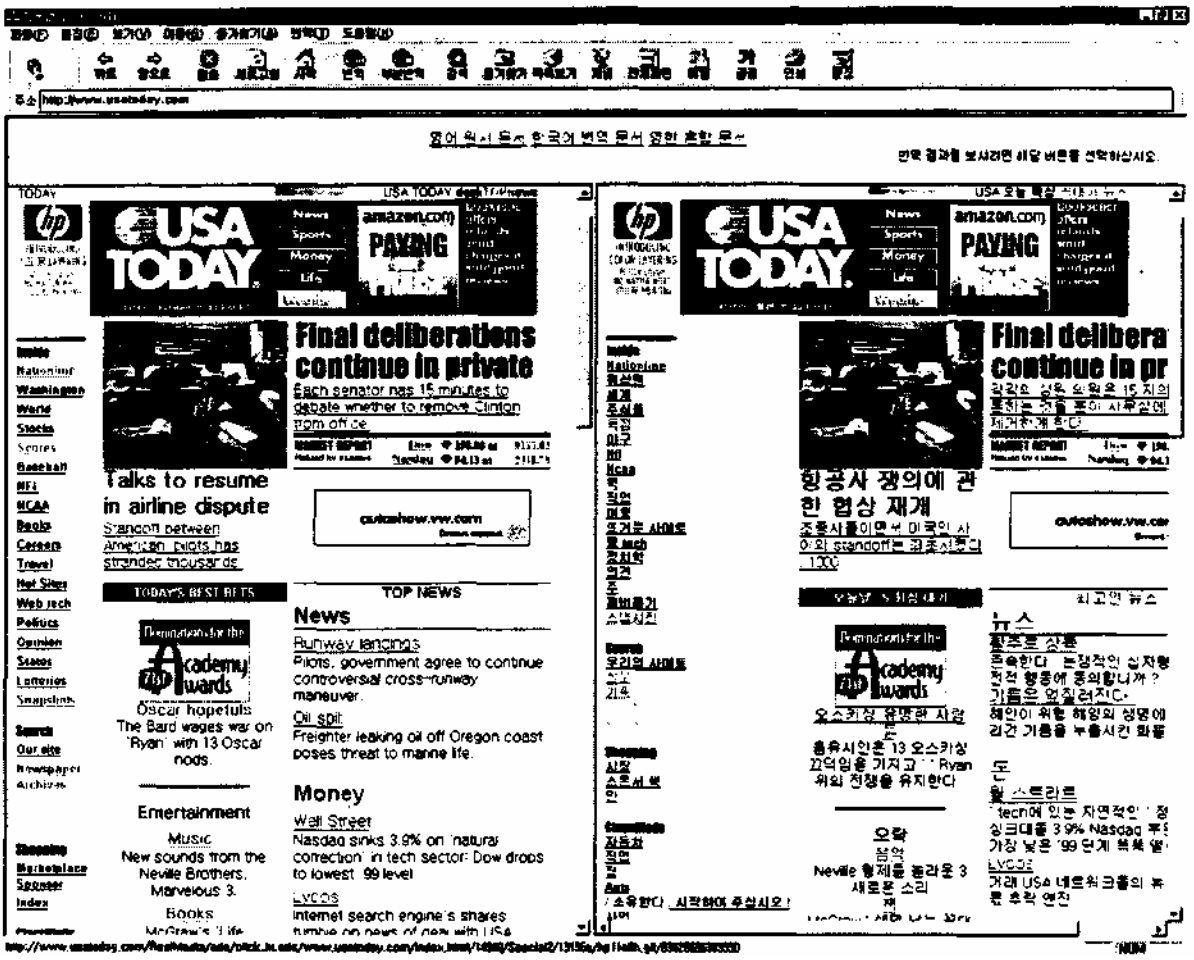
Kupiec, J. (1992). Robust part-of-speech tagging Using a Hidden Markov Model, Computer Speech and Language, pp. 225-242.

Ratnaparkhi, A. (1996). A Maximum Entropy Model for Part-Of-Speech Tagging, in Proceeding of Conf. on Empirical Methods in Natural Language Processing.

Schmid, H. (1994). Part-of-Speech Tagging with Neural Networks, Int. Conf. on Computational

Linguistics, pp. 172-176.

Yuh Sanghwa, Hanmin Jung, and Jungyun Seo. (1999). Neutag : A Hybrid Neural Network English Tagger with Pre-Fail Softener. ICCPOL99.



[Figure 4] Example of Translation by English-to-Korean Web Translator "FromTo/Web-EK"