What is Needed the Most in MT-Supported Paper Writing

Chang Hyun Kim, Oh-Woog Kwon, Young Kil Kim

ETRI, NLP Team
161 Gajeong-dong, Yuseong-gu, 305-350 Daejeon, Korea
{chkim,ohwoog,kimyk}@etri.re.kr

Abstract. This paper addresses our system which provides an effective method to write an English paper suitable for international conferences and analyze the system's pros and cons through the user's data collected from operating the system for 6 months. The system consists of Korean-English paper MT module supported by user interaction environment. Our original Korean-English paper MT system was quite useful for understanding, but not satisfactory for writing. So, we analyzed our system to trace what caused such dissatisfaction. We classified the analysis results into three main categories, that is, the errors in the source sentence itself, the errors of our MT system, and the absence of the appropriate domain-specific expression information. For each category we provide an alternative method and show the effectiveness through analyzing the user's data. We can confirm that our system can be used quite usefully for paper writing.

Keywords: Machine Translation, User Interaction, Korean-English Translation, Paper writing

1. Introduction

Many Koreans who are not fluent in English have difficulty in writing a scientific paper or technical documents in English. Although the state-of-the-art Korean-English MT system is quite useful for understanding, many people hesitate to use the MT system for paper writing. Understanding does not necessarily require perfect sentences, but writing papers does require impeccable grammars and correct, native expressions.

The main purpose of the original Korean-English paper MT system (Kim, 2007) was to help researchers or students to submit their papers to a conference or an academic journal. This system had been developed through customization of the patent MT system (Hong, 2005), which is currently serviced by KIPO (Korean Intellectual Property Office) and is being used by more than 20 countries with positive feedbacks from foreign users. The customization process includes construction of translation resources specialized in scientific papers, modification of the engine to reflect the linguistic characteristics of academic papers. Moreover, a Controlled-Language (CL) guided Korean rewriting checker is provided to correct the errors in Korean spelling and sentence structure. Language model component reports the unlikely or unnatural English expression.

Several beta testers of the original MT system reported that it was very helpful in writing a paper, but that was not enough. They said that the user interface was inconvenient, and they wanted to understand what caused the mistranslations, and how to correct them. Besides, the MT output still contained erroneous expressions even the users rewrite sentences according to the guidelines of the CL-checker.

We analyzed those reports and found 3 main reasons: the errors in the source sentence itself, the errors of our MT system, and the absence of the appropriate domain-specific expression information.

In this paper, we provide alternative methods to cope with those problems within our user interaction environment. As a result, authors can interact with the system through modification of source sentences, correction of engine errors, and correction of target sentence expression.

In section 2 we will survey some major works on controlled language and interactive MT. Section 3 deals with the three steps of user interaction process in detail. At each subsection, the simulation of the user interaction will be described with proper examples. We implemented our system and opened the beta site to users for 6 months. In Section 4, we show the statistics we got from operating our system and analysis result. Finally, conclusions and future work are presented in section 5.

2. Related Works

Re-designing the traditional MT system for the improvement of the translation quality can be driven from two perspectives: Firstly, a controlled language can be adopted to enhance the readability and translatability. Secondly, an interactive MT system can be implemented to collect meta-information through user interactions to resolve the ambiguities and errors from the translation process. There is no clear definition as to what a controlled language or the interactive MT system should be like.

A controlled language has usually a restricted vocabulary and syntax rules. Most of the works on a controlled language focus on how to design a grammar rules and lexicon for a given language (Mitamura, 1999; Adriaens & Schreuers, 1992; Fuchs et al, 1999). The emphasis of major controlling could be put on the lexicon (AECMA, 1995) or on the syntax restrictions (Lehrndorfer, 1996). In our current setting, the major controlling takes place on the syntactic level because small set of syntactic restrictions affects the performance seriously. To split a long sentence into a fragment of simple sentences which are controlled by our scheme, we used a set of syntactic rules which has lexical/grammatical features. (Shirai et al., 1998) reports the improvement of translation quality by 20% through applying rewriting rules to Japanese to English translation.

The interactive MT system provides UI functions connected with the engine which includes a translation model and a language model that are used to produce the translation candidates. The target sentence under construction serves as the medium of communication between an MT system and its user (Foster et al., 1997, Langlais et al., 2000). In such an environment, human translators interact with a translation system that acts as an assistance tool and dynamically provides a list of translation candidates. To extend a type of translation models, a hybrid approach was suggested (Yamabana, 1997).

The language model that is adopted at the end of our MT system has been widely used as a post-processing step to enhance the generation performance in MT systems (Liu et al., 2003).

3. Interactive Machine Translation System

The design principles for our system are as follows; maximization of user's engine control, user's optional control, provision of full information about error correction, and user-friendly interface.

Maximization of user's engine control means that users get full control on the intermediate process of the translation, for example, the modification of morphological/syntactic analysis and target word selection. So, if a user wants to check and modify the intermediate results of the engine in the course of translation, engine errors can be corrected and more improved translation is possible.

To provide the function of engine control is one thing and to use it is another. User's optional control means that users can control the process of the translation engine as much as they want and can turn off functions which they don't want. If a user is poor in English, he/she probably wants to focus only on the rewriting of the Korean sentence. If a user knows the translation process well and wants better translation, he/she is going to revise errors from translation engine more deeply. A user can select the level of engine control and the system provides only such items that a user has selected.

Provision of full information means that our MT system provides full information that is related to the improvement of translation to the user. Those are morphological/syntactic analysis results,

the translation result, link information between Korean and English words, error candidates in Korean and English sentences. The system also offers information on what the information means exactly and how to handle error candidates effectively by providing examples.
To implement user-friendly interface, the system detects both user's action and the environment and determines what the user wants in such environment. The information is represented as easy and instinctive as possible. The user can see the effect of correction by pressing the translation button right away.
Figure 1 shows the main window of our system. It consists of four sub-windows, that is, Korean window, English window, working window, and sentence structure window. The Korean window on top left shows the Korean sentences to be translated. The English window on bottom left shows the translated English sentences. The modified English sentence by the user is also saved in English window. The working window on top right shows one Korean sentence and the corresponding English sentence which is the user's current concern. The sentence structure window shows the syntactic structure of the Korean sentence on working window. Basically, one node of the tree is a simple sentence which is linked with the corresponding English translation. Link information on simple sentence level is more easy to grasp the structure and find errors than on word level. Word-level syntactic structure can be seen also if a user clicks the '+' on the tree.

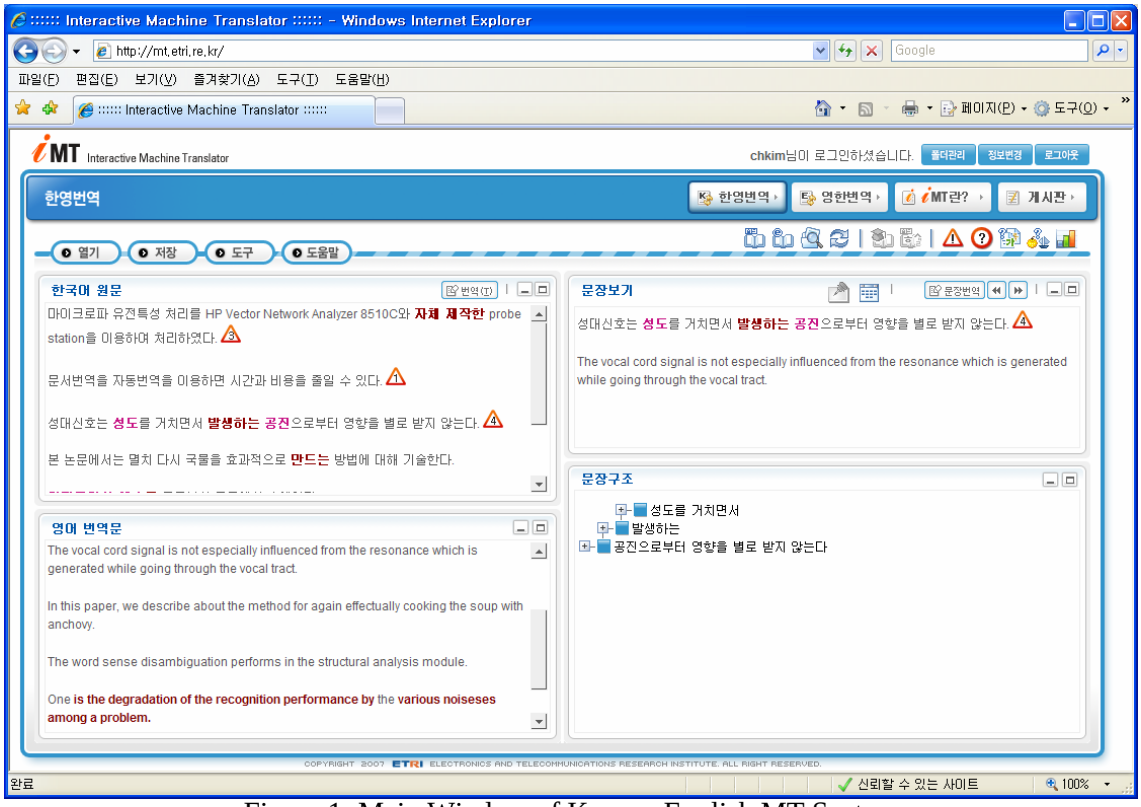

Figure 1: Main Window of Korean-English MT System

3.1 Korean Sentence Modification

Korean sentences are scanned and analyzed by using morphological, morpho-syntactic, syntactic information and candidates for modification are reported to the user. Modification candidates include both error correction candidates and quality improvement candidates. Most of the errors in Korean sentences are spelling errors and spacing errors which must be corrected before translation. Such error candidates are reported to the user through triangle marks as in Figure 2. If a user presses the triangle button, the error-related part of the sentence are

highlighted and at the same time an information box is popped up which describes the error type and how to handle it.

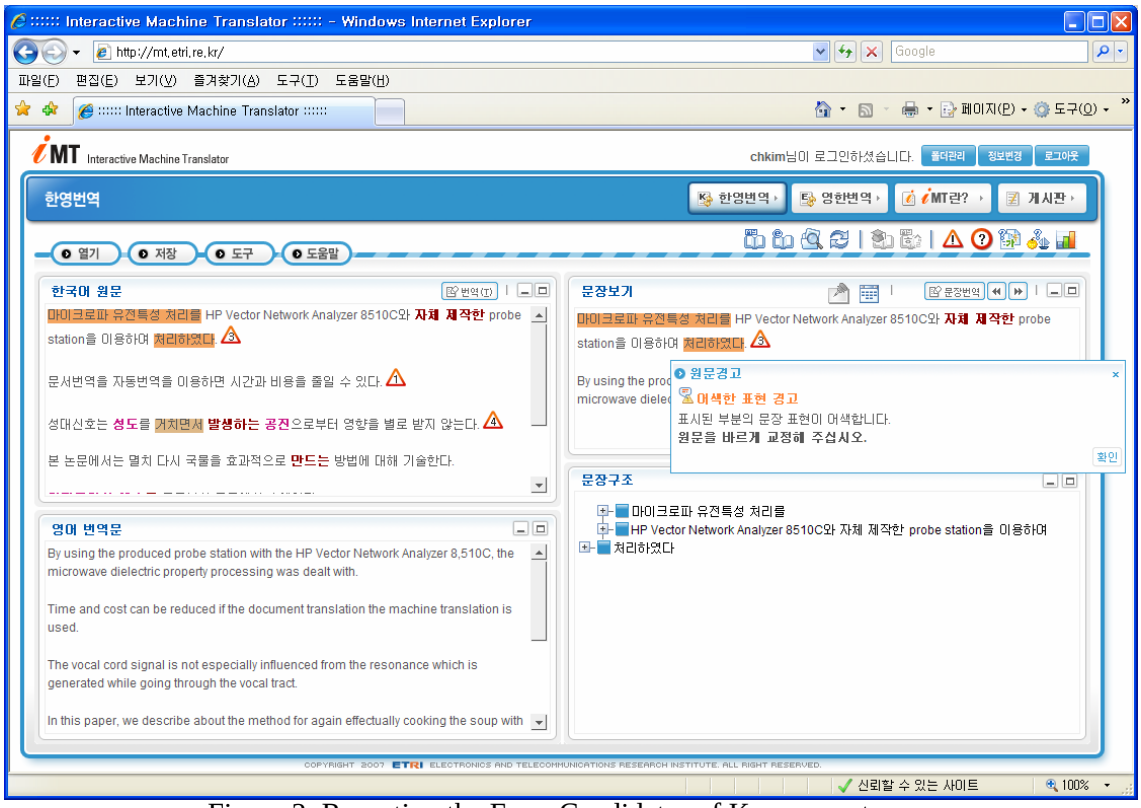

Figure 2: Reporting the Error Candidates of Korean sentences

Rewriting the Korean sentence needs to be done both for translatability and readability. But, sometimes they conflict with each other where we prefer translatability. For example, the appropriate use of auxiliary postpositions in Korean can enhance the readability for human in many cases, but it sometimes causes errors in translation. But, most of the time, improvement in readability leads to improvement in translatability. Ambiguous words or long sentences often make mistakes for human, much more for machine translation. Examples of ambiguous words are as follows.

(a) 기본적인 HMM 모델**도** 사용하는 경우
(b) 기본적인 HMM 모델**에서** 벗어나지 않고
(c) 얼굴 검출을 **할** 경우에는
(d) 지문의 방향영상을 **구할** 경우

Auxiliary-postpositions cause case ambiguities. In (a), '도' has case ambiguities between subject/object/adverb case, and a user is asked about whether '도' can be replaced by case-postpositions such as '이(subject)','를(object)' or others. If it is better not to modify, then no action is needed. Case-postpositions can cause ambiguities also. In general, '에서' has several meanings and can be replaced by other less ambiguous words for each meaning. In (b), the better alternative of '에서' is '로부터' and the original translation 'deviate in the basic HMM

model' is changed to 'deviate from the basic HMM model'. '하다'(which means do) is one of the most frequently used verb in Korean and the abuse of '하다' often leads to deterioration in translatability and even in readability. So, if '하다' is considered to be better to modify, '하다' is reported for modification as in (c). The modified sentence "얼굴을 검출할 경우에는" has the translation 'if the face is detected' instead of the original translation 'if the face detection is done'. Verbs acting like pro-verb also causes ambiguities as '구하다' in (d). The user is asked about whether to change '구하다' into '계산하다(compute)', '얻다(get)', or '구하다(save)'. Modifications on the structure are as follows.

(e) … **형상을** … 여러 **형상을** 다단계 모델의 구조로 생성하는 기술을 말한다.

Unlike English, there exist double subject/object phenomena in Korean, the translation of which is various depending on their semantic characteristics. But, many double subject/object sentences are erroneous in reality. (e) is such an example. So, double subject/object sentences with the possibility of error are reported to the user.
In Korean, ellipses are frequently occurred in various ways as the following.

(f) 첫번째 프레임에서 **얼굴** 검출하는 경우
(g) 성능 개선을 수행하는 경우
(h) 오류를 **검출**, 수정하는 과정에서

The ellipsis of postposition and obligatory case as in (f) is easy to detect and the user needs to change '얼굴' into '얼굴을' . Unlike English, subject ellipsis is common in Korean and if a Korean transitive verb has no subject in a sentence, the user is asked about whether to convert it into intransitive or not. The English translation of a Korean transitive verb requires a subject all the time. The intransitive version of (g) is '성능 개선이 수행되는 경우' and the translation doesn't need subject. On the contrary, as in (h), the ellipsis of suffix part in a light verb is not easy to detect and the failure of the detection leads to the wrong syntactic analysis and wrong translation. The unabridged form of '검출(detection, noun)' is '검출하다' (detect, verb) in (h), where the verb '검출하다' is mis-interpreted as noun '검출'(detection). For this kind of ellipsis we use lexical co-occurrence information and also syntactic patterns. Lexical co-occurrence dictionary has entries like '오류-를-검출하다'.
In addition to the fore-mentioned, there are still other kinds of problems in translation as in the following :

(i) 증가를 가져오다
(j) 이렇게 하여 나오는 정보는
(k) 최대수는 3*n* 이며, 최소수는 *n* 이 된다

Although the Korean expression is natural to the native Korean, the translation can be awkward in many cases. For example, the translation of (i) is 'bring increment'. The correct translation is 'increase' which is the translation of '증가시키다'. Non-informative expressions can lead the mis-translation also. For example, the translation of (j) is 'information which does in this way

and come out' where '하다' is obsolete. The modified sentence '이렇게 나오는 정보들은' get the , translation 'information coming out in this way'. The application of agreement/concord can improve the translation quality also. The translation of sentence (k) is 'The maximum number is 3*n* and the minimum number becomes *n*' which is very faithful to the source sentence. With respect to the standpoint of agreement/concord, '이 된다' can be modified into '이다' in (k) and the translation is 'The maximum number is 3*n* and the minimum number is *n*'. Generally, human doesn't want to repeat the same vocabulary in writing. But, the application of agreement/concord and therefore the use of the same vocabulary is a very good way for machine translation.

The modifications described in this section are obtained automatically or semi-automatically through corpus analysis and they are still needed to be complemented.

3.2 Engine Error Correction

Engine errors are not easy for a user to understand and correct. So, items reported to users are needed to be understandable and manageable. We only report such errors like morphological, syntactic analysis errors and word translation errors to the user.

The morphological errors are part-of-speech tagging errors and segmentation errors which can be found indirectly through scanning the translation result. These errors can be modified by correcting the morphological analysis directly. If a user presses the morphological analysis button, the morphological analysis result is popped up. Figure 3 is the morphological analysis result of "본 논문에서는 멸치 다시 국물을 효과적으로 만드는 방법에 대해 기술한다.". Here, '다시' is wrongly tagged as adverb, the user can fix it.

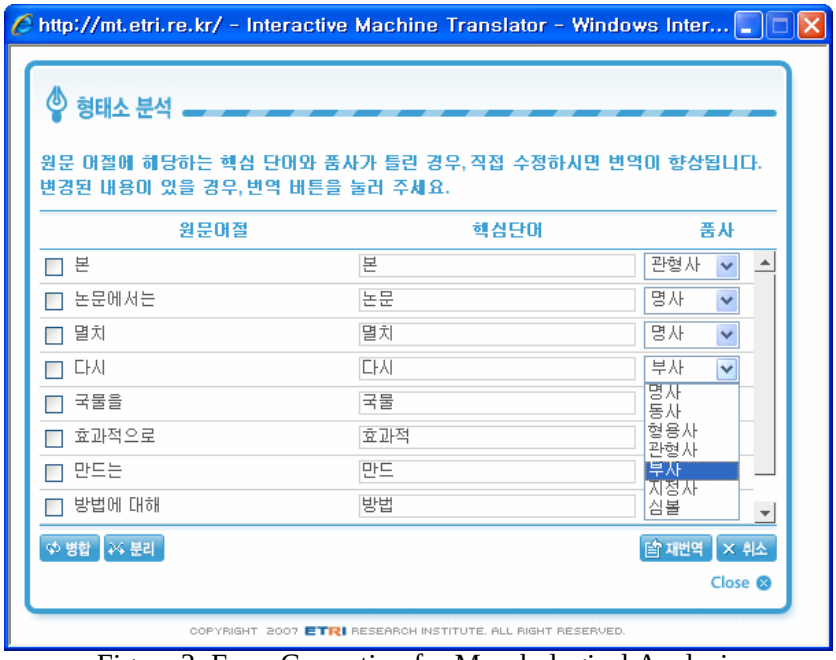

Figure 3: Error Correction for Morphological Analysis

Syntactic analysis result is displayed on the sentence structure window. Each line in a tree corresponds to a simple sentence and its translation is linked with its translation. Figure 4 is the original and modified tree. Dtra&drop is used for structure modification.

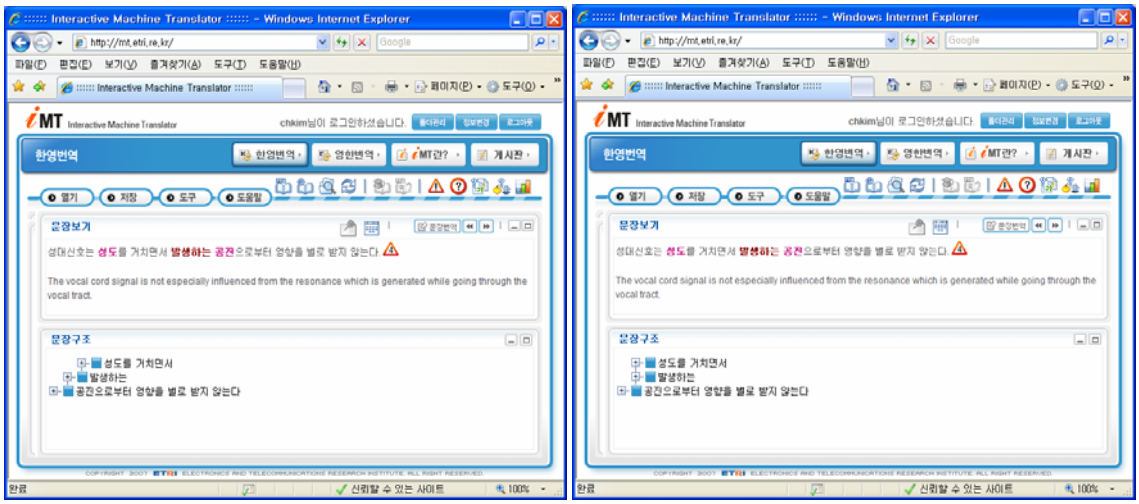

Figure 4: Sentence Structure before and after the Correction

Word translation errors can be modified by pressing the suspicious word and selecting the right one among several candidates or by typing in the right one directly.

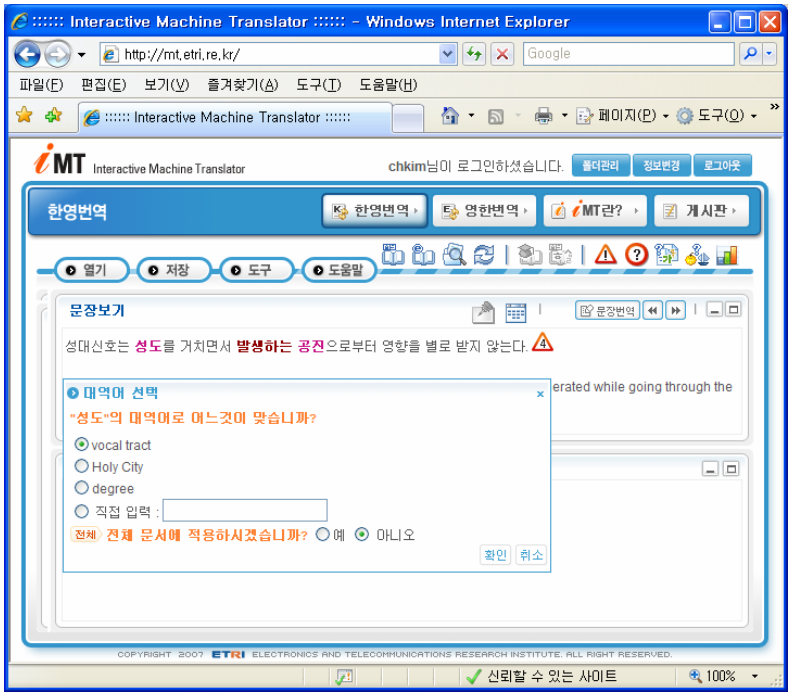

Figure 5: Word Translation Error Correction

3.3 Target Sentence Correction

Source sentence modification and engine error correction can improve the translation, however it still may not be satisfying. This is because our paper MT system is pattern-based system. Our MT system generates target sentences mainly based on pattern resources such as sentence patterns, verbal patterns, noun patterns and etc. When the wrong patterns are matched and used in generation, the translated English sentences may contain erroneous expressions. Even when the patterns are correctly matched, the English counterpart may contain somewhat unnatural translation. For this reason, we employed the language model for post-processing. For example, the translation of “필요성이 대두되고 있다” is “a necessity is occurring”. The system reports to the user that “a necessity is occurring” is scarcely used and shows all possible English

translation for “필요성이 대두되고 있다” by consulting the dictionary for each Korean word and combining the candidates. In this case, those Korean words ‘필요성’(necessity) and ‘대두되다’(occur, come to the front, raise, show itself, be raised) are consulted and combined. Each combination expression is retrieved from the English paper database for its examples. From this information, the user gets a hint on how to correct the expression. Sometimes some expressions look unnatural to the user even though the system regards them as natural. For the user’s confidence, the system provides the function to retrieve the same expression as the one in the translation as in Figure 6.

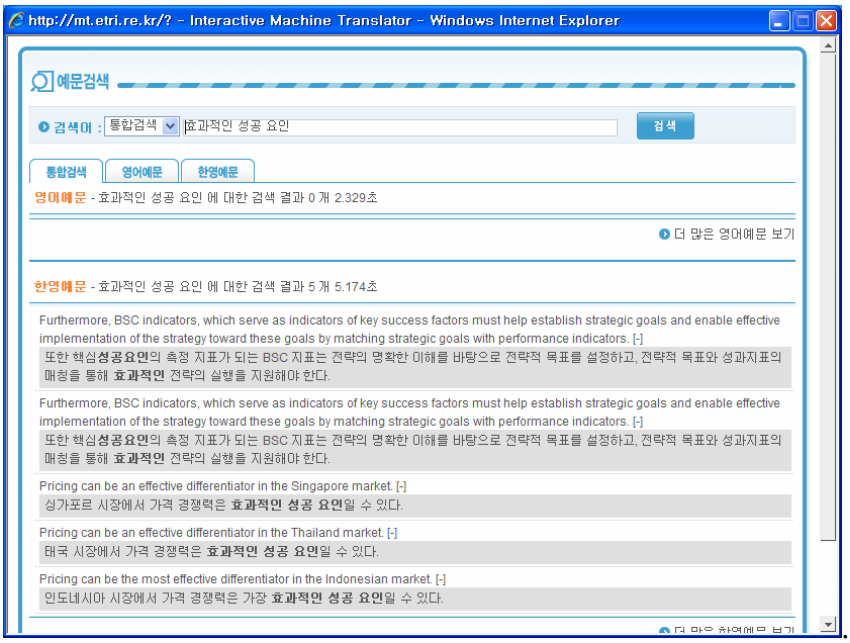

Figure 6: Expression Search

4. Evaluation

We have implemented our system and opened it to users for 6 months. The main users are researchers and university students in the field of science and technology.

Total # of Users	Total Length of Stay(day)	Average(stay/user)
694	1,827	2.633

The total number of users and their length of stay are as the above. It is hard to say that our system is useful for paper writing according to the above statistics. Some few can write an English paper just in 2.6 days but most of people can’t.

Total # of Users	Total # of Sentences	Average (Sent./user)
694	39,506	56.925

The average sentence translated per user is as the above. It is still not so promising but a little bit more positive than the previous one. We can think that if a user thinks our system is useful, then the user is sure to use more complex function of the system in addition to the simple translation such as the source sentence modification, the engine error correction or the target sentence modification. These kind of functions entail the re-translation(RT).

# of RT Users	Total # of RT Sentences	Average (RT/user)

344	7,142	20.762

The retranslation statistics are as the above. This is somewhat different from the previous ones. We can think that around 50% of the users consider our system is interesting. Let's narrow our interest to the 344 users(focus user(FU)) and renew the above statistics.

# of FU	Length of Stay	Average Stay	# of FU Sent.	Average Sent.
344	1,442	4.192	38,096	110.744

The average length of stay per FU is 4.192 and the average number of sentence for translation is 110.744 for paper translation. These statistics surely tells that our system can be contributed to some of FU's needs although we do not know their needs exactly. Then let's analyze our system more deeply for those needs. The below is the details for RT sentences.

# of FU Sentences	Total # of RT Sentences(ratio)	# of Non-RT Sentences(ratio)
38,096	7,142(18.75%)	30,954(81.25%)

The above says that FU users use 81.25% of the total FU sentences as it is without modifying anything at all. If we analyse this statistics on the assumption that our system is useful, 81.25% means that the most preferred function of the system is the translation function itself. It can also be said that the users are either satisfied with the translation result or although they are not satisfied with the translation quality, they know the limits of the system and use the translation result usefully according their needs by any means. In fact, phrase-level or simple-sentence level translation is quite correct.

Total # of RT	MorphErr	TreeErr	CaseErr	SenseErr
7,142	37	47	236	97

The above is the statistics for complex functions. MorphErr is the number of RT triggered by morphological errors, TreeErr by structural analysis, CaseErr by case analysis error, SenseErr by sense disambiguation errors. The other sentences except the above 4 cases are the case for the Korean sentence correction. Korean is an agglutinative language and has some somewhat complicated spelling and spacing rules. So Koreans have often difficulties in writing Korean sentences accurately. So, errors in Korean sentences cause poor translation results. By just correcting the Korean sentences, the performance of the system can be improved by a large margin. The most frequently used functions among the above are CaseErr and SenseErr. The reason seems to be that these two functions are quite intuitive and simple for use and also are easy to see the effect of correction directly. On the other hand, the functions MorphErr and TreeErr are somewhat difficult for the layman to understand and use and also the effect of the correction may not be easy to identify directly sometimes.
The post-editing function is not dealt with in this section. In reality we don't have the statistics for the function.

5. Conclusion

In this paper, we presented the Korean-English paper machine translation system allowing the user interaction and evaluated the implemented system by opening it to users for 6 months. To obtain high quality translation we redesigned our MT system and applied the new design principles: maximization of user's engine control, user's optional control, provision of sufficient information about error correction, and user-friendly interface.

The evaluation statistics show that almost half of the users consider our system can be of any help to their needs for English paper writing. 81% of the translation results are used as it is without any modification, which means that the performance of our system can be said satisfactory to some extent. A Korean sentence is somewhat complicated to write and includes errors often. This fact is identified by the RT statistics and just correcting the Korean sentence errors can enhance the degree of satisfaction to the translation quality. Case disambiguation and sense disambiguation functions are frequently used that are easy to understand and see the effects directly. Morphological analysis and structural analysis functions are rarely used partly because of the difficulties in understanding and identifying the effects directly.

From the evaluation we can conclude that the basic translation quality is the most important in paper writing MT systems but, the function of just correcting the source sentence errors can be enormous help. In addition to that, simple functions such as case disambiguation and sense disambiguation can be very helpful, but user are not so interested in complex functions such as morphological analysis and syntactic analysis.

In the future, we will continually improve the translation performance of our MT translation engine and improve the user interaction based on the analysis performed in this paper.

References

AECMA : A Guide for the Preparation of Aircraft Maintenance Documentation in the International Aerosace Maintenance Language, AECMA Simplified English , 1995.

Adriaens, G. and D. Schreuers. 1992. From COGRAM to ALCOGRAM: Toward a controlled Eng-lish grammar checker, *COLING* , 595-601.

Foster, G, Pierre Isabelle and Pierre Plamondon. 1997. Target-Text Mediated Interactive Machine Translation, *Machine Translation*.

Fuchs, N. E., U. Schwertel and R. Schwitter. 1999. Attempto Controlled English (ACE) Language Man-ual, Version 3.0, Technical Report, Department of Computer Science, University of Zurich.

Hong, M., Y. Kim, C. Kim, S. Yang, Y. Seo, C. Ryu and S. Park. 2005. Customizing a Korean-English MT System for Patent Translation, *MT-Summit*.

Kim, Y., M. Hong and S. Park. 2007. CL Guided Korean-English MT system for scientific papers, *CICLing*

Langlais, P., G. Foster, and G. Lapalme. 2000. TransType : a computer-aided translation typing system. *In Workshop on Embedded Machine Translation Systems*.

Lehrndorfer, Anne. 1996. Kontrolliertes Deutsch, Gunter Narr Verlag, Tuebingen.

Liu, Fu-Hua, Liang Gu, Yuqing Gao and Michael Picheny. 2003. Use of Statistical N-gram Models in Natural Language Generation for Machine Translation. *MT-Summit*.

Mitamura, Teruko. 1999. Controlled language for multilingual MT, *MT-Summit*.

Roh, Y., Y. Seo, K. Lee and S. Choi. 2001. Long Sentence Partitioning using Structure Analysis for Machine Translation, *NLPRS*.

Shirai, S., S. Ikekaha, A. Yokoo and Y. Ooyama. 1998. Automatic Rewriting Method for Internal Expressions in Japanese to English MT and Its Effects, *In proceedings of the Second International Workshop on Controlled Language Applications(CLAW98)*.

Seo Y, K. Lee and S. Park. 2001. CaptionEye/EK: English-to-Korean Caption Translation System using the Sentence Pattern, *MT-Summit*.

Yamabana, K., S. Kamei, K. Muraki, S. Doi, S. Tamura and K. Satoh. 1997. A Hybrid Approach to Interactive Machine Translation – Integrating Rule-based, Corpus-based, and Example-based Method, *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence.*