Getting Professional Translation through User Interaction

Young-Ae Seo, Chang-Hyun Kim, Seong-Il Yang, Young-gil Kim

Natural Language Processing Team
Electronics and Telecommunications Research Institute (ETRI)
161 Gajeong-dong, Yuseong-gu, Daejeon, 305-700
South Korea
{yaseo, chkim, siyang, kimyk}@etri.re.kr

Abstract

This paper addresses an effective method to write an English paper suitable for international conferences using our Korean-English paper MT system supported by efficient user interaction environment. Our original Korean-English paper MT system is quite useful for understanding, but not for writing. We analyzed the problem of our system and found 3 main reasons, 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. In this paper, we provide an effective method for each problem within our user interaction environment. Representative sentence error patterns are obtained through large amount of paper corpus analysis and the user is reported on those kinds of errors for modification. Error candidates of the MT system are reported to the user and the corrections from the user are feedbacked to the system. Finally, the system detects English expressions with low frequency and also proposes more suitable domain-specific expression candidates. The final translation sentences we can get from our system shows 93.3 % accuracy, which, we think, is almost as the level suitable for conference submission.

1 Introduction

Many Koreans who are not fluent in English writing feel difficulties in writing a scientific paper or technical documents in English. While the current performance of state-of-the-art Korean-English MT system is very useful for understanding, Korean paper authors still hesitate to use MT systems to write English papers because writing a paper needs more precise expressions. Understanding the meaning of sentences does not require perfect sentences which have impeccable grammar and correct expressions, but writing an official document does.

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 by customizing the patent MT system (Hong, 2005), which is currently serviced by KIPO (Korean Intellectual Property Office) and used by more than 20 countries with positive feedbacks from foreign users. The customization process included a construction of translation resources specialized in scientific papers, and the modification of engine modules after linguistic studies of academic papers. Moreover, to overcome the obstacles for "professional" translations, a Controlled Language (CL) guided Korean rewriting checker to avoid the linguistic obstacles that may affect the translation accuracy and a language model module to present the candidates of unnatural expression to a paper author were implemented.

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 did not understand why wrong-translated sentences were generated and how to correct them because the system did not provide sufficient information on error correction. Besides, the MT output still contained erroneous expressions however users rewrite sentences according to the guideline of the CL-checker.

We analyzed problems 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 an effective method for each problem within our user interaction environment. [1] Korean authors can interact with system in three methods, that is, source sentence modification, engine error correction, target sentence correction.

In section 2 we 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 show the experimental results in section 4. Finally, conclusions and future work are presented in section 5.

2 Related Works

To maximize the translation quality, redesigning the traditional MT system can be driven from two perspectives: Firstly, a controlled language can be adopted to enhance the readability and transibility. Secondly, the interactive MT system can be implemented to collect the meta-information from user interactions, so that it can avoid the ambiguity and errors which are produced 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). It was critical to be

[1]This work was supported by the IT R&D program of MIC/IITA. [2006-S-037-02, Domain Customized Machine Translation Technology Development for Korean, Chinese, English]

balanced by whether the emphasis of major controlling should be put on the lexicon (AECMA, 1995) or on the syntax restrictions (Lehrndorfer, 1996). In our current setting, the emphasis of controlling takes place on the syntactic level because small set of syntactic restrictions affects the performance more 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. In a similar case (Shirai et al., 1998) of applying rewriting rules to Japanese to English translation, the translation quality is improved by 20%.

The interactive MT system provides UI functions supporting the engine which includes a translation model and a language model 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. User Interaction with MT System

The design principles of our MT system are as follows: maximization of user's engine control, user's optional control, provision of sufficient information about error correction, and user friendly interface.

Maximization of user's engine control means that users can get control of the full process of the translation engine, for example, the error correction in morphological/syntactic analysis and target word selection. We concluded that if users cannot control the full translation process of the engine, they may not get high-quality translation result. This is why we give users the right of maximal control.

While users get the right of maximal control, they also have the right of choosing control level. User's optional control means that users can control the process of the translation engine as much as they want to do. If a user is relatively poor in English, he/she may put emphasis on the rewriting of the Korean sentence. If he/she wants to get professional translation quality, he/she is going to revise all the errors from the engine. The level of engine control can be set by the user.

For provision of sufficient information to fix the errors generated by the engine, the MT system provides the morphological/syntactic analysis result and the generation result to users, and informs where the errors are suspected in Korean and English sentences. The system also offers users how to handle these errors by providing correction-related information.

To implement user friendly interface, the system detects user's action and presents the appropriate action. The system also reflects user's correction directly to the translation result. Whenever user changes translation-related information, the system feedbacks the corrected information to the engine and regenerates English sentence in real-time.

Figure 1 shows the main window of the MT system, which contains four sub-windows, that is source sentence window, target sentence window, translation result window, and sentence structure window. The source sentence window shows the Korean sentences to be translated. The target sentence window shows the translated English sentences. The revised English sentence by user is also reflected in target sentence window. The translation result window shows one Korean sentence and the corresponding translation result which a user is currently concerned in. The sentence structure window shows the analysis information about Korean sentence in the translation result window and the corresponding English sentence in the simple sentence unit. A user can edit Korean or English sentence and correct the translation engine's errors through four sub-windows and the editing result is directly reflected in all sub-windows.

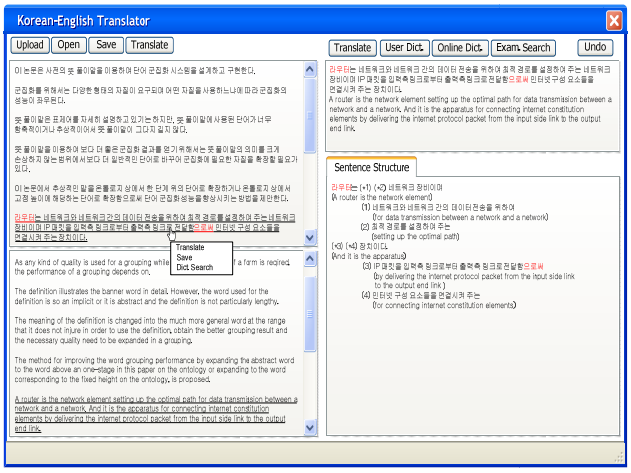

Figure 1: Main Window of Korean-English MT System

Users can interact with our MT system through three steps; source sentence modification, engine error correction and target sentence correction. In this chapter, we will describe these in detail.

3.1 Source Sentence Modification

Source sentences are scanned first 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. Errors in a sentence are mainly spelling errors and spacing errors. But, there can be too many such error candidates in a sentence and we decided not to report them directly but indirectly through link information between a Korean word and its English word.

As in Figure 2, if a user points a Korean/English word, their corresponding words are highlighted at the same time in all windows. Therefore, if a user finds an unexpected translated word while scanning the result, he/she can know on the spot where mis-translation came from. Modification can be done at any window and the

modified results are reflected in all three windows at the same time.

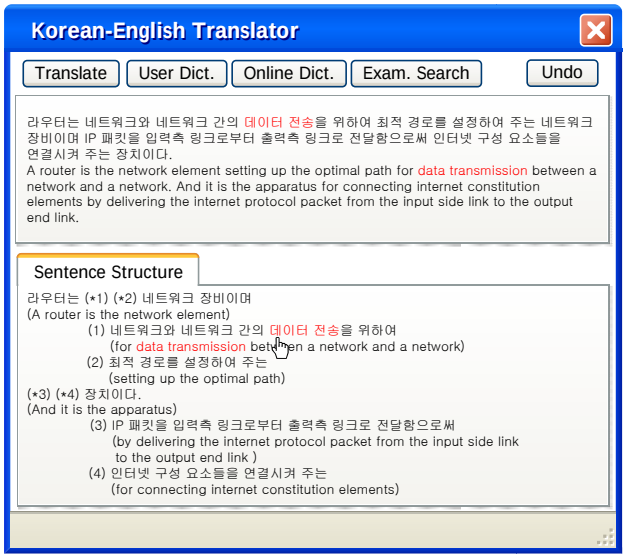

Figure 2: Indirect Reporting of Error Candidates

Unlike error candidates, quality improvement candidates are reported directly. Quality improvement includes modifications both for translatability and readability. But, if a modification conflicts between translatability and readability, translatability is preferred. For example, the appropriate use of auxiliary postpositions can enhance the readability for human in many cases, but it is not the case for translatability. So, modifications on ambiguous words are mainly for translatability. Modifications for readability are in most cases effective for translatability as in the case of sentence length modification. A too long sentence is not easy for a reader to understand and also is not easy to translate.

The most basic but effective modification among others is on sentence length and use of comma. If a sentence violates a given condition on sentence length and the use of comma[2], it is reported directly to the user as in Figure 3. In Figure 3, the proposed sentence breaking positions are highlighted on all windows. Here, the sentence structure window shows the overall sentence structure and makes it easy for a user to evaluate the sentence. A user can split the sentence by just editing the text on any window and the result is reflected on all windows.

Modifications on 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

[2] For example, if a sentence is over 20 words with 3 or more predicates, the system proposes sentence breaking position candidates.

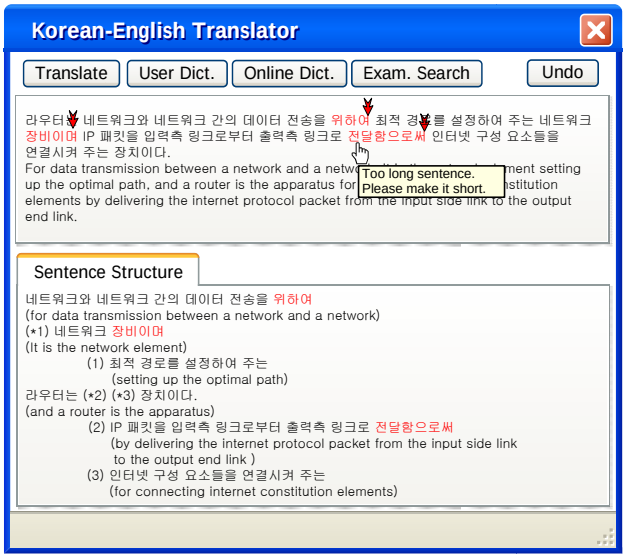

Figure 3: Sentence length and the use of comma

Auxiliary-postpositions cause case ambiguities. In (a), '도' has case ambiguities between subjective/objective/adverbial case, and a user is asked about whether '도' can be replaced by other postpositions such as '이(subjective case)', '를(objective case)',etc.. If it is better not to modify, then no action is needed. The system remembers the user's action including no-action and does not report the same information later if not asked. Case-postpositions can cause ambiguities also. In general, '에서' can be replaced by '이(subjective case)', '로부터(adverbial case)', etc. and in (b), '로부터' is the better expression. By modifying '에서' to '로부터', the original translation 'deviate in the basic HMM model' is retranslated to 'deviate from the basic HMM model'. The report of modification information is context-dependent. If '에서' is determined to be appropriate for example, it is not reported to the user. '하다(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 the conditions for the modification of '하다' are satisfied, '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' to '계산하다(compute)', '얻다(get)', '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. In addition to that, many double subject/object sentences are erroneous in reality. (e) is such an example. So, double

subject/object sentences which are suspicious of errors 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), (g) is easily detected and it is reported to the user for the restoration of the omitted element. A transitive verb can be converted to intransitive verb and in that case the omitted subject is not needed to be restored in many cases. So, if the verb is transitive and the omitted case is the subject, the user is also asked about whether to convert the sentence into intransitive sentence or not as in (g). The modified version of (g) is '성능 개선이 수행되는 경우' and the translation doesn't need the omitted subject in the original sentence. On the contrary, 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 translation. The appropriate form of '검출(detection, noun)' in (h) is '검출하다(detect, verb)'. But the system fails to detect it and, the verb '검출하다' is mis-interpreted as noun '검출'. For the detection of this kind of ellipsis, we currently 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 modifications, there are other kinds of modifications.

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

Although the Korean expression looks natural, the translation can be awkward in many cases. For example, the translation of (i) is 'bring increment', which is somewhat unnatural. The natural translation is 'increase' and it is obtained by modifying the source sentence into '증가시키다(increase)'. Additional expressions in Korean which are not informative at all can lead the translation to the wrong way. For example, the translation of (j) is 'information which does in this way and come out'. From the translation we can decide that '하다' is obsolete in this sentence. By modifying the sentence into '이렇게 나오는 정보들은', we can get the translation 'information coming out in this way'. The application of agreement/concord in a sentence can improve the translation quality also. The sentence (k) looks very natural, but the translation 'The maximum number is 3n and the minimum number becomes n' is somewhat unnatural. But, the translation is very faithful to the source sentence. If we introduce agreement/concord in a sentence, we can modify '이 된다(become)' into '이다(be)' and get the translation 'The maximum number is 3n 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. So, engine error 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 of complex nouns. In a sentence '나는 새를 보았다', for example, if the noun '나(I)' is wrongly tagged as verb, the user can detect it by scanning the translation as explained in previous section and modify it by using the right button on the mouse. The right button shows context-dependent action. The wrong segmentation of complex noun entails wrong translation. Segmentation errors are also modified through the right button.
Syntactic analysis result is displayed on the sentence structure window. Each line corresponds to a simple sentence and its translation is also displayed. The whole Korean sentence can be re-constructed by traversing the structure down from the top and by traversing the first-encountered sentence index first. The forefront simple sentence is the head sentence and the indented backside simple sentence is the dependent simple sentence. By just scanning the forefront simple sentence, the user can verify whether the structure is the same as expected or not. Figure 4 is the system result with wrong syntactic structure and Figure 5 is the modified right one. Structure modification is done in both ways using drag&drop function or using the right button on the mouse.

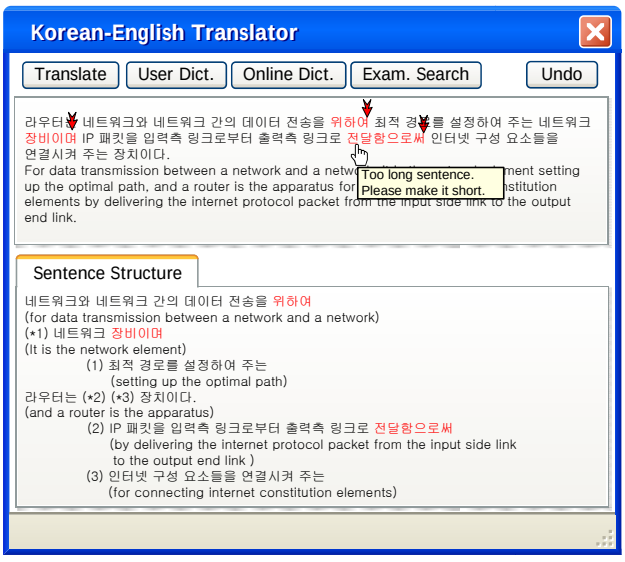

Figure 4: Sentence Structure before Correction

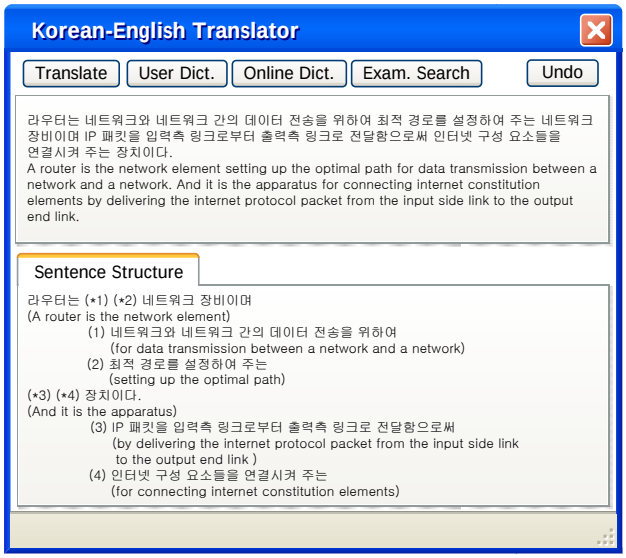

Figure 5: Sentence Structure after Correction

Word translation errors can be modified by selecting the right one among several candidates or by typing in the right one directly. Unknown words are always reported for its translation. Figure 6 shows an example.

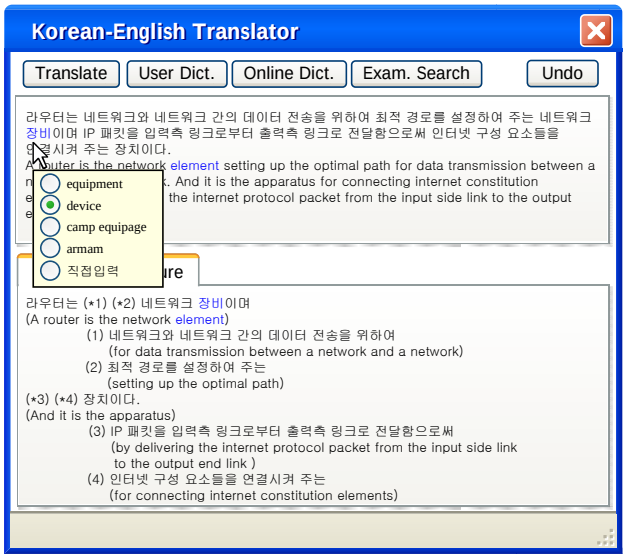

Figure 6: Word Translation Error Correction

3.3 Target Sentence Correction

The user can get more improved English sentences through the source sentence modification and engine error correction. However, he/she still could not satisfy the translation quality. One of the main reasons is that our paper MT system is pattern-based system. The system generates target sentences mainly based on pattern resources such as sentence patterns, verbal patterns, noun patterns and etc. When the wrong patterns are used in generation, the translated English sentences may contain erroneous expressions. For this reason, we have employed the language model module for the proof-reading of the system's translation (Kim, 2007).

Even though the system reports awkward English expression candidates to the user by computing the probability of the translated English word sequence, the user may not know how to modify them into natural expressions. If the user knows English well, he/she can correct the awkward part based on one's own linguistic knowledge. But, if not, he/she should depend on a Korean-English dictionary and search example expressions. When the exactly matched example expression is found in the dictionary or in the example corpus, the translation quality will be improve. But even in this case, the process is time-consuming. If it is not the case, most users will try to combine target words in the dictionary and search the combined expressions in Google or some other places. Through this time-consuming process, users can barely get the right expressions. It can be effective, but troublesome and repetitive work. So, our system tries to support this process more conveniently. That is, our system provides information on awkward expression candidates, dictionary lookup, and example search simultaneously.

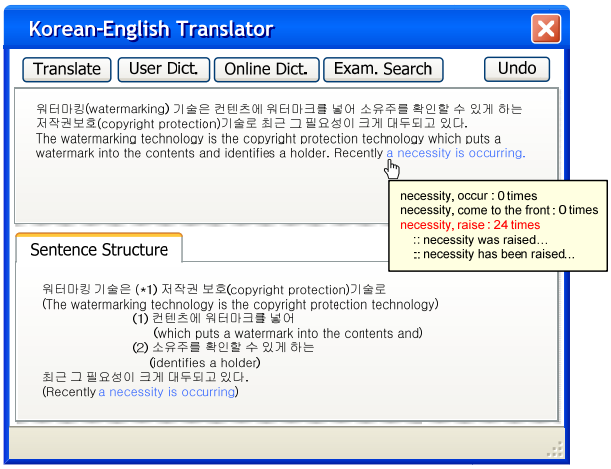

Figure 7: Correct Expression Candidates Search Result

Figure 7 shows the system's search result on correct expression candidates. The system detects the awkward English expression 'a necessity is occurring' based on the language model, and reports to the user by blue-colouring the expression. When the user puts the mouse point on the expression, the system provides all possible English target words by using Korean dictionary. The possible English candidates for '필요성' is necessity, and the candidates for '대두되다' are 'occur', 'come to the front', 'raise', 'show itself', 'be raised'. Then, it generates all possible combination among target words and searches the each expression from our own English paper database. Finally it displays the search result with frequency information as in Figure 7. From this information, the user gets a hint on how to correct the awkward expression, so he/she can change the wrong expression 'a necessity is occurring' into the right expression 'necessity has been raised.'

Sometimes natural expressions may look unnatural to a user. In this case, for the user's confidence, the system can provide example sentences with the same example expressions by searching our English scientific paper database as in Figure 8.

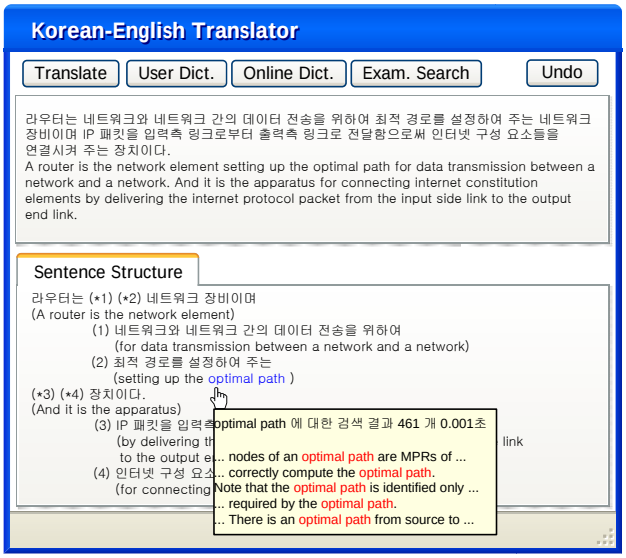

Figure 8: Example Expression Search Result

4. Evaluation

We evaluated the translation accuracy with 100 test sentences randomly extracted from paper corpus. The average length of test sentences was 18.54 eojeols[3], which was normalized in order to reflect the length of the real paper sentences.

Because the goal of the evaluation was to see not only the translation quality of finally generated sentences but also how much improve the translation quality by three steps of correction, we translated the test sentences by 4 manners. The first translation was conducted only by the translation engine without user's interaction. The second was generated just by source sentence modification, or CL-guided Korean rewriting. The third translated sentences were made by source sentence modification and engine error correction. The Korean morphological/syntactic error and target word generation error were targets of the engine error correction. The final translation was obtained through the full correction process including the target sentence correction based on the engine-provided dictionary and example expression search.

Table 1 describes the scoring criteria for evaluating translation accuracy. The translation accuracy (TA) is calculated by the formula, TA = $[(S_1 + S_2 + \ldots S_n)/n]*(100/4)$ (%) where S_1 is the evaluated score of the first sentence and "n" is the number of evaluated sentences (Kim, 2007).

Score	Criterion
4	The meaning of a sentence is perfectly conveyed
3.5	The meaning of a sentence is almost perfectly conveyed except for some minor errors (e.g. wrong article, stylistic errors)
3	The meaning of a sentence is almost conveyed (e.g. some errors in target word selection)
2.5	A simple sentence in a complex sentence is correctly translated
2	A sentence is translated phrase-wise
1	Only some words are translated
0	No translation

Table 1: Scoring criteria for translation accuracy

Two PhD candidates of Korea University of Science and Technology took part in translating test sentences and two professional translators were hired for assessing the accuracy. The estimated scores were summed and the average was taken as the accuracy. The accuracy of four manners was shown in Table 2.

Translation Method	Translation Accuracy
Paper Machine Translation Engine	71.38% (285.5/400)
Source Sentence Modification	79.25% (317/400)
Engine Error Correction	85.13% (340.25/400)
Target Sentence Correction	93.25% (373/400)

Table 2: Accuracy according to translation manners

The accuracy of the untouched translation result is 71.38%. As shown in the translation accuracy, the almost original translation sentences convey their meanings but were not enough to submit to an international conference. The improvement of 7.89% was caused by the source sentence modification, which is higher improvement than that of engine error correction, 5.88%. This is because many morphological and syntactic engine errors were revised by the source sentence modification. The large portion of the 5.88% improvement was caused by the correction of the target word error, especially of Jargon. The user's target sentence correction based on the example search improved about 8.12%. The final version of translation sentences could directly submit to the conference without major corrections of English by native speakers or whatsoever.

5. Conclusion

In this paper, we presented the Korean-English paper machine translation system allowing the user interaction.

To obtain professional translation, we redesigned the original paper MT system with holding the new design principles: maximization of user's engine control, user's optional control, provision of sufficient information for error correction, and user friendly interface.

The evaluation showed that the translation accuracy can be improved by about 21.9% through the user actions such as the source sentence modification guided by CL checker, correction of engine error generated through the process of morphological/syntactic analysis and target word generation, and target sentence correction guided by the language model and engine-providing example expressions. The translated sentences with the accuracy of 93.25% were in the state to be directly submitted to the conference without major corrections.

[3] An eojeol is a spacing unit corresponding to a bunsetsu in Japanese.

In the future, we will continually improve the translation performance of our MT engine. And, we will introduce the dependency language model to capture the long-distance dependency.

References

AECMA : A Guide for the Preparation of Aircraft Maintenance Documentation in the Interna-tional Aerosace Maintenance Language, AECMA Simplified English (1995).

G. Adriaens and D. Schreuers : From COGRAM to ALCOGRAM: Toward a controlled Eng-lish grammar checker, in COLING [COL92], (1992) 595-601.

George Foster, Pierre Isabelle, Pierre Plamondon : Target-Text Mediated Interactive Machine Translation, Machine Translation (1997)

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

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

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

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

Anne Lehrndorfer : Kontrolliertes Deutsch, Gunter Narr Verlag, Tuebingen (1996)

Fu-Hua Liu, Liang Gu, Yuqing Gao and Michael Picheny : Use of Statistical N-gram Models in Natural Language Generation for Machine Translation (2003)

Teruko Mitamura, Controlled language for multilingual MT, MT-Summit (1999)

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

Shirai, S., Ikekaha, S., Yokoo, A. and Ooyama, Y : Automatic Rewriting Method for Internal Expressions in Japanese to English MT and Its Effects, In proceedings of the Second Interna-tional Workshop on Controlled Language Applications(CLAW-98) (1998).

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

Yamabana K., Kamei S., Muraki K., Doi S., Tamura S. and Satoh K. : 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 (1997)