Controlled Korean for Korean-English MT

Munpyo Hong [a], **Chang-Hyun Kim**[b]

[a] Department of German Lang. & Lit.
Sungkyunkwan University
53 Myeongnyun-dong
Seoul, Korea
skkhmp@skku.edu

[b] Natural Language Processing Team
ETRI
Gajeong-dong 16
Taejeon, Korea
chkim@etri.re.kr

Abstract. This paper addresses the issues in designing the so-called 'Controlled Korean' for a Korean-English MT system. Controlled Language is a sublanguage of a natural language which is supposed to improve the readability and the translatability of a text. Much effort has been made to design a controlled language for major international languages such as English, German, Spanish and etc. However, little effort has been made yet to design a Controlled Korean in the context of machine translation. In this paper we introduce the concept of the Controlled Korean we have developed for a Korean-English MT system and compare the Controlled Korean with the Controlled English and Controlled German from the perspective of the translatability. The result of our experiments shows that in designing a Controlled Language, not only the linguistic characteristics of the language but also the characteristics of an MT-System must be taken into account.

Keywords: Controlled Language, Controlled Korean, Machine Translation, Translatability, Negative Translatability Indicators

1. Introduction

Controlled Language is a sublanguage of a natural language designed to improve the readability and the translatability of a text. Originally, the concept of a controlled language was introduced in the field of technical documentations to prevent the misunderstanding of texts. [1] In the last couple of years the necessity of a controlled language has increased significantly, especially in the technical documentation domain, as the number of pages to be translated has increased enormously.

The increase of the volume of the documents to be translated made it necessary to employ a full automatic translation system like an MT or a machine-aided translation system like translation memory. In Korea, however, MT systems have not been widely welcomed by the experts in the localization business, because the quality of the translations failed to match their expectations.

The recent researches on the controlled languages show that the use of a controlled language can generally lead to the improvement of machine translation quality, thus reducing the cost of

[1] cf. Esselink(2000), Göpferich(1995)

post-editing. The general assumption behind the idea of the controlled language is that the cost of the use of a controlled language that can be paraphrased as 'pre-editing' is lower than that of 'post-editing' in using an MT system for localization.

Previous works on the controlled language in the context of MT have mainly focused on the impact of the controlled language on MT. Aikawa et al. (2007), for example, addressed the impact of the controlled English for MSR-MT. Lehrndorfer (1996), Lehrndorfer&Schachtl (1998) dealt with the controlled German for German-English MT.

In this paper, we will introduce the concept of the 'Controlled Korean' for Korean-English MT. We will not only show the impact of the Controlled Korean on Korean-English MT, but also share our experience in designing the Controlled Korean.

2. Controlled Korean

2.1.Background

Since early 2000, MT has been paid much attention in Korea because of the fast growth of the Internet. However, most of the efforts to bring out the off-the-shelf products into the market have failed mainly due to the difficulty in Korean syntactic parsing and the failure in the domain adaptation.[2] One of the few Korean-English MT systems actively used not only in Korea but also in foreign countries is the Korean-English Patent MT system developed at ETRI.[3] The performance of the patent MT system is good enough for foreign patent examiners to retrieve the patent documents of their interest in English.

The quality of the Korean-English patent MT is good enough for cross-language information retrieval, however not good enough, if it is to be used for other purposes like academic paper authoring. If an MT system is to be used for academic paper writing in English, the quality of the translation must be far superior to that of the patent MT system, the purpose of which is not to produce a perfect translation, but to produce a translation just enough for the understanding. However, we are very well aware of the difficulties and obstacles in improving the performance of an MT system in a short period of time. One way to improve the translation quality is to employ a controlled language for MT. Kim et al. (2007) showed that even some simple writing rules can improve the translation quality for around 10 %.

2.2.Designing Controlled Korean

To design a controlled language for MT, the purpose of the application must be properly understood. Most of the controlled languages introduced so far have without exceptions limited lexicons and writing rules. This is possible and useful, when the controlled language and the MT system are used by a homogeneous group of users like employees in a company. However, in the academic paper authoring setting, the scenario should be somewhat different, i.e., the users have all the different interests of their own. Thus, it is almost impossible to enforce them to use only the allowed lexicon and not to use any words they like. We therefore gave up introducing the restricted Korean lexicon and rather focussed on the writing rules which impact on the translation quality most.

The philosophy in designing the Controlled Korean was to give the authors as much freedom as possible. The authors write an English academic paper that will be automatically checked against Controlled Korean writing rules by a Controlled Korean Checker.[4] Thus we investigated on the translation errors that might be occurred by stylistic errors that are hard for an MT system to treat properly.

To do this, we manually checked 40,000 Korean-English machine translation pairs. We scored all the translations from 0(poor) to 4(very good). The translations above 3 points(good) were

[2] Korean-Japanese MT, however, is widely used in both countries
[3] Hong et al. (2005)
[4] cf. Kim et al. (2007) about the Controlled Korean Checker module of IMT system

excluded. We assumed that the syntactic analysis and the transfer of those sentences succeeded without fatal errors. Then we focused on the translations that were scored below 3. Many of them failed in the morphological or syntactic analysis for some reasons. We were interested in those cases where, though the morphological and the syntactic analysis succeeded, but the translation was poor. In such cases generally the so-called 'Konglish (English in Korean style)' was produced. We assumed that the reason why such 'Konglish' was produced was that the source Korean sentence was poor.

We found out 8 most frequent writing errors that Korean authors commit and that have the most significant effects on the performance of an MT system.[5]

Table 1: Most frequent writing errors that effect on MT

Error Types	**Description**
Subject-Predicate mismatch	Semantic mismatches between the subject and the predicates in a sentence
Topic Markers in the Sentence Initial Position	Sentence initial NPs with topic markers are underspecified w.r.t. their case
Ambiguous expressions	Use of specific ambiguous words such as 'ulo', 'hata', 'tayhata' and etc. [6]
Spoken language expressions	Use of spoken language type expressions
Double subject/object construction	Though these constructions are legitimate, they are difficult to analyze correctly
Punctuation	If a sentence is long, the correct punctuation helps the anlaysis
Minor grammatical errors	Though the grammatical errors are not supposed to be treated by a controlled language checker, some grammatical errors are so frequent and not conceived by the authors that they have to be checked before the syntactic analysis
Light-verb expressions	Frequent use of unknown light-verb contructions that can be substituted by a single verb

Another important issue in designing a controlled language is the learnability of the controlled language. Lehrndorfer(1996), for example, criticized the AECMA Simplified English for its poor learnability. Especially, in the academic paper authoring setting, it will be very difficult to train the users with the Controlled Korean. Therefore, the application of the above rules must be automated with a Controlled Korean Checker. However, the formalization of above rules is not always simple. For example, the semantic mismatch of the subject and the predicates in a sentence is very difficult to detect automatically. It would not be possible without deep semantic processing. However, the deep semantic processing technique is currently not available. Therefore we collected lexical clues with which we can detect the semantic mismatches on the surface level. Currently about 6,000 manually constructed lexical rules and metarules are employed in the Controlled Korean Checker.

[5] Strictly speaking, some of these 'errors' are not a grammatical error, but we stick to the term 'error' for the simplicity of the expression

[6] 'ulo' corresponds roughly to English 'as' and 'with'

3. Experiment

3.1.The impact of Controlled Korean on MT

O'Brien(2005) introduced the concept of the 'negative translatability indicators'. The negative translatability indicators are the specific linguistic constructions or phenomena that negatively effect on the quality of MT. We performed an experiment to find out the negative translatability indicators in Korean-English MT. We collected 50 sentences for each error type in the table 1. Each sentence was rewritten only as the Controlled Korean Checker suggests, i.e., even if the sentence contained the error of the given type, if it is not detected by our Controlled Korean Checker, the sentence was not corrected. Table 2 shows the result of our experiment.

Table 2: Negative Translatability Indicators in Korean-English MT

Error Types	**Improvements**
Ambiguous expressions	+25%
Subject-Predicate mismatch	+21%
Double subject/object construction	+20%
Topic Markers in the Sentence Initial Position	+12.5%
Punctuation	+12.5%
Spoken language expressions	+7.5%
Minor grammatical errors	+4.2%
Light-verb expressions	+4.2%

It turned out that the 'ambiguous expressions', 'subject-predicate mismatch', and 'double subject/object construction' error types are the most important negative translatability indicators. Most of these error types were related to resolving ambiguities in the analysis. As mentioned, in order to detect the subject-predicate semantic mismatch, many rules on the lexical level must be written. In other words, the more lexical rules we have, the better is the the performance of the Controlled Korean Checker, hence Korean-English MT system expected to be.

The overall performance of the Korean-English MT system supported by the Controlled Korean Checker was improved for 4.25%. As Kim et al.(2007) showed, if we can provide the checker with more rules, we can improve the performance of the MT system for more than 10%.

3.2.Controlled English/German

We were interested not only in finding out what are the negative translatability indicators in Korean, but also in learning if the MT paradigm plays a role in designing a controlled language. For this purpose, we conducted another experiment. In this experiment, we analyzed Controlled English by Aikawa et al.(2007) and Controlled German proposed by Lehrndorfer(1996).

Aikawa et al.(2007) introduced the MS Controlled English designed for the MSR-MT System which is a statistics-based MT system. They showed that Controlled English not only improves the performance of a rule-based MT system but also a statistics-based MT system. The following table shows the negative translatability indicators for 4 language pairs.

Table 3: Negative Translatability Indicators in Aikawa et a.(2007)

	Eng.=>Ara.	**Eng.=>Chin.**	**Eng.=>Fr.**	**Eng.=>Du.**
1	Formal Style	Formal Style	Short Ambiguous Sentences	Formal Style
2	Hyphens	Attachment	Formal Style	Capitalization

3	Short Ambiguous Sentences	-ing clauses	Spelling	Spelling
4	Capitalization	Spelling	Adjective/Verb Ambiguity	Short Ambiguous Sentences
5	Spelling	Long Sentences	Capitalization	Long Sentences

Contrary to our expectation, in 3 out of 4 language pairs, the 'Formal Style'[7] was the most influential negative translatability indicator. In case of Controlled Korean, most of the indicators were related to resolving the ambiguities in the source language.

In the experiment with the Controlled German, we employed a rule-based German-English MT system. We were interested in learning what can be the negative translatability indicators in German-English rule-based MT. In the preparation stage, we collected 250 sentences from German technical documents in the IT domain. On the next step, we re-wrote the sentence according to the guidelines proposed by Lehrndorfer(1996). Among 250 sentences 64 sentences were re-written. By applying the Controlled German, we could improve the translation accuracy of the whole sentences for 2.41%. If we consider only those sentences that were re-written with Controlled German, the translation accuracy rose for 9.28%. The following table shows the negative translatability indicators in rule-based German-English MT.

Table 4: Negative Translatability Indicators German-English rule-based MT

	Controlled German Writing Rules
1	Don't use relative sentences
2	Don't use long sentences
3	Follow the coordination construction rules
4	Put the subject in the sentence initial position
5	Don't omit connectives

In the experiment with Controlled German, the rules resolving the ambiguities are the most important negative translatability indicators. Though more comprehensive experiments should follow to back up our conclusion, we assume that in designing a controlled language for MT, the paradigm of the MT should be well taken into account.

4. Conclusion

In this paper we proposed 'Controlled Korean' for Korean-English MT. We also shared our experience in designing a controlled language. We didn't introduce the controlled lexicon as done in many other researches, but we focused on the writing rules. As the learnability is a very important factor for the acceptability by the users, we tried to formalize the Controlled Korean rules as much as possible. As a result, we could improve the translation accuracy of the Korean-English MT system backed up by a Controlled Korean Checker for 4.25%.

Another important issue in designing a controlled language for MT is whether the MT paradigm should be considered. Our experiment and the comparison of our result with Aikawa et al.(2007) showed that the MT paradigm could play a very important role in designing a controlled language for MT. Our experiment with Controlled German for rule-based MT showed different results from that of Aikawa et al.(2007). For a rule-based MT, ambiguities-resolving rules seem to play important roles, whereas for a statistics-based MT, rather the rules

[7] 'Formal Style' relates to the use of spoken language expressions or slangs.

that normalize the source expressions so that the decoding can be performed more easily seem to play more important roles.

References

AECMA (1995): AECMA Simplified English, A Guideline for the Preparation of Aircraft Maintenance Documentation in the International Aerospace Maintenance Language, Issue I

Aikawa et al. (2007): “ Impact of Controlled Language on Translation Quality and Post-editing in a Statistical Machine Translation System” , Proceedings of MT-Summit XI, 1-8

Esselink, B. (2000): A Practical Guide to Localization, John Benjamins Publishing Company, Amsterdam/Philadelphia

Göpferich, S. (1995): Textsorten in Naturwissenschaft und Technik. Pragmatische Typologie - Kontrastierung -Translation. Tüubingen: Narr.

Hong, M., Kim, Y., Kim, C., Yang, S., Seo, Y., Ryu, C. & S. Park (2005): Customizing a Korean-English MT System for Patent Translation, Proceedings of MT-Summit X, 181-187

Kim, Y., Hong, M. & S. Park (2007): CL-Guided Korean-English MT System for Scientific Papers, Lecture Notes in Computer Science, vol.4394, 409-419, Springer Verlag

Lehrndorfer, A. (1996): Kontrolliertes Deutsch: linguistische und sprachpsychologische Leitlinien für eine (maschinell) kontrollierte Sprache in der technischen Dokumentation. Tüubingen: Narr.

Lehrndorfer, A. / R. Mangold (1997): "How to Save Money in Translation Cost", TC-Forum 97-2, URL: http://www.techwriter.de/tc-forum/pdf/editions/ tcf972s.pdf

Lehrndorfer, A. / S. Schachtl (1998): Controlled Siemens Documentary German and TopTrans, TC-Forum 98-3, URL: http://www.tc-forum.org/topictr/tr9contr.htm

Ley, M. (2005): Kontrollierte Textstrukturen. Ein (linguistisches) Informationsmodell für die Technische Kommunikation. Dissertation, Justus-Liebig-Universitäat Gießsen.

Mitamura, T. / Nyberg, E. H. (1995): “ Controlled English for Knowledge-Based MT: Experience with the KANT System” , Proceedings of TMI-95.

Mitamura, T. (1999): “ Controlled Language for Multilingual Machine Translation” . Proceedings of MT-Summit 1999

Möller, M. (2003): Grammatical Metaphor, Controlled Language and Machine Translation, Proceedings of EAMT/CLAW 2003

Nübel, R. (2004): "Evaluation and Adaptation of a Specialised Language Checking Tool for Nonspecialised Machine Translation and Non-expert MT Users for Multi-lingual Telecooperation", Proceedings of LREC 2004

O'Brien, S (2005): Methodologies for Measuring the Correlations between Post-Editing Effort and Machine Translatability, Machine Translation, vol.19, no.1, 37-58

O'Brien, S (2006): Pauses as Indicators of Cognitive Effort in Post-editing Machine Translation Output, Across Languages and Cultures vol.7, no.1, 1-21

O'Brien, S. & J. Roturier (2007): How Portable are Controlled Language Rules? A Comparison of Two Empirical MT Studies, Proceedings of MT Summit XI, 345-352

Reuther, U. (2003), "Two in one - can it work? Readability and translatability by means of controlled language". Proceedings of the 4th International Workshop on Controlled Language Applications, Dublin, Ireland.