Toward Practical Use of Machine Translation

Hitoshi Isahara

Toyohashi University of Technology
isahara@tut.jp

Abstract. This paper presents an overview of our research activities on machine translation conducted at Toyohashi University of Technology. I focus on how to make machine translation useful for real world business, not mentioning quality improvement of translation engines. I present here three approaches for making machine translation practical; i.e. simplifying the Japanese source text, extracting and listing salient expressions and their equivalents in a document and enhancing the post-editing process. This study is important from both a business perspective and an academic perspective.

Keywords: Machine Translation, Simplified Language, Term Extraction, Post-editing.

1 Introduction

Various services, such as information retrieval and information extraction, using natural language processing technologies trained by huge corpora have become available. In the field of machine translation (MT), corpus-based machine translations, such as statistical machine translation (SMT) and example-based machine translation (EBMT), are typical applications of using such volumes of data in real business situations. Thanks to such huge available data, current machine translation system is enough high quality for some specific language pairs. But still some people have doubt about usefulness of machine translation, especially for translation among different types of languages, such as Japanese and English. One study examined for what types of people current machine translation systems are useful [1], by simulating the retrieval and reading of web pages in a language different from one's mother tongue. However, there has been little research to verify the technologies which make MT systems more useful in real world situations.

In this paper we focus on how to make machine translation useful for real world business. I present here three approaches; i.e. simplifying the Japanese source text, extracting and listing salient expressions and their equivalents in a document and enhancing the post-editing process. This study is important from both a business perspective and an academic perspective.

H. Isahara and K. Kanzaki (Eds.): JapTAL 2012, LNAI 7614, pp. 23–27, 2012.

2 Problems of Translation between Japanese and English

Developers of Japanese-to-English and English-to-Japanese machine translation systems face more difficulties than counterparts providing systems translating, for example, English-to-French. This is because Japanese is very different in syntax and semantics from English, so we often need some context to translate Japanese into English (and English into Japanese) accurately. English uses a subject-verb-object word order, while in Japanese, the verb comes at the end of the sentence, i.e. a subject-object-verb order. This means that we have to provide much more example sentence pairs of Japanese and English compared to when translating most European languages into English, as they also use a subject-verb-object order. The computational power required for Japanese to come up with accurate matches is enormous. And accuracy is particularly necessary for businesses selling their products overseas, which is the reason why it is needed to help Japanese companies provide better translated manuals for their products.

Faced with such obstacles, we conduct researches on quality improvement of MT engines, which include five-year national project on development of Japanese-Chinese machine translation system [2]. In parallel with this kind of MT research, we are taking a three-step approach to improve the MT quality in real life environment: simplifying the Japanese source text (controlled language), enriching lexicon and enhancing the post-editing process (Figure 1).

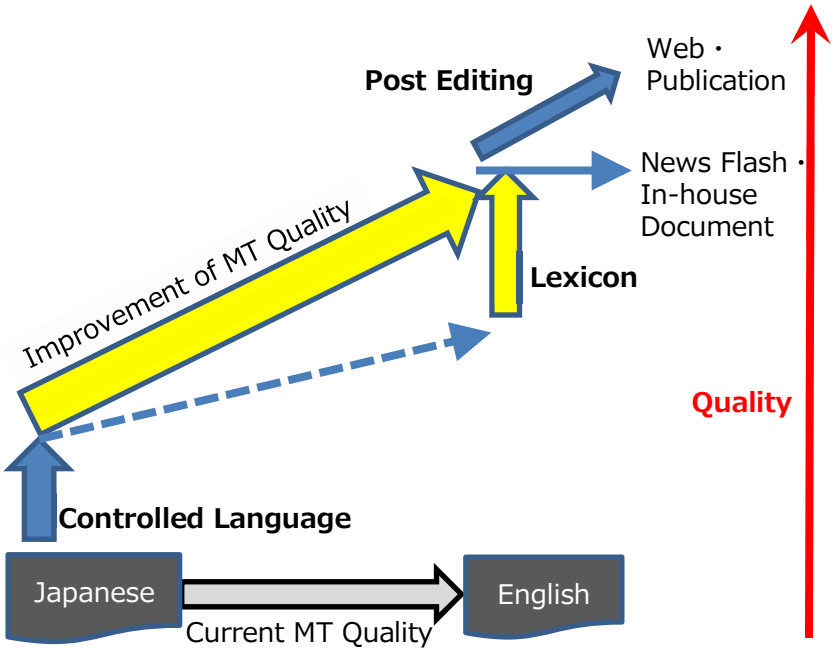

Fig. 1. Quality of Translation

For controlled language, e.g., simplified Japanese, we are devising a set of guidelines and rules for writers of Japanese manuals that will be used as the source for MT. These rules include writing shorter and simpler sentences; adding the subject when missing; and providing context when there is ambiguity. As for enriching lexicon, we are extracting and listing salient expressions and their equivalents in a document and store them into translation dictionary. For post-editing, which can be costly and time-consuming, we are conducting an experiment using foreign students of Toyohashi University of Technology to post-edit the MT output of English version of the university's Web site into their own languages.

Combining above mentioned techniques, MT will be practically useful tools for real world translation. Our approach, then, is not to focus on just one aspect of MT. Rather we want to improve and support the entire machine translation process.

3 Simplified Japanese [3,4]

Output quality of MT heavily depends on the quality of analysis of input sentences. Long and complex sentences in syntax and semantics are mostly very difficult for automatic analyzers to output proper structures. Therefore, restricting structures of input text will be beneficial for MT system to achieve high quality translation. We seek to address this challenge by investigating the feasibility of developing a 'controlled Japanese' with explicit restrictions on vocabulary, syntax and style adequate for authoring technical documentation. This research is being conducting in collaboration with an automobile related company in Japan.

We aimed to build translation awareness within a global Japanese company where non-professional authors are called upon to write 'global job manuals' for internal dissemination. Following an analysis of current practice, we devised a document template and simple writing rules which we tested experimentally with MT systems. Sentences violating the rules were extracted from the original data and rewritten in accordance with the respective rule. The original and rewritten sentences were then translated by MT systems, and the inputs and outputs were submitted to human evaluation. Overall, native-speaker judges found that the quality of the Japanese was maintained or improved, while the impact on the raw English translations varied according to MT system. Then, we explained our template and rules to employees of the company and asked them to write their manuals articulating the know-how using our template and rules. We are currently investigating their documents and trying to identify the most promising avenues for further development.

One of the other possibilities of controlled (or simplified) language is translation between two languages both of which are properly controlled. If we train SMT with parallel controlled language corpus, the SMT can translate controlled input into controlled output with high quality. Some of multilingual MT systems are combination of MT engines for two languages and translations between non-English languages are performed via English. Such cascade translation usually amplifies errors during translation. Using controlled English as a pivot would be promising solution of this problem.

4 Extracting and Listing Salient Expressions and Their Equivalents in a Document [5]

Quality of translation, especially its informativeness for human readers, is affected by whether technical expressions are translated properly or not. We are trying to compile automatically parallel term dictionary using documents in a specific domain.

There are several methods to acquire new words from large amount of text and some of them showed high performance for compound nouns. Our aim is to acquire technical terms which include not only compound nouns but also longer phrases such as "Extraction of Informative Expressions from Domain Specific Documents" in Japanese. The method uses morpheme based n-gram to save processing time and space, therefore the acquired terms are compounds of one or plural number of morphemes.

Because we use morpheme strings as input, each morpheme in the string is usually not an unknown words but stored in the dictionary of morphological analyzer. We extract compound nouns and longer phrases which are new terms as a whole. Or, we can say we extract salient terms including compound nouns and noun phrases which are written in Japanese but may contain many English words.

Our term acquisition method consists of two stages: an extraction of candidate terms ("Candidate Selection") and a guess as to terms ("Unithood Checking"). First, the statistical indicators we defined are used to select all one-morpheme to ten-morpheme strings that appear at least once in a large number of documents, and also appear repeatedly in several documents. In this way, we have enabled a computer to emulate a human sense to recognize and understand unknown terms. Next, the strength of connection between the constituent morphemes of each candidate term is assessed to arrive at a guess as to whether or not it is in fact a term.

We extracted salient terms both in Japanese and English from parallel documents, e.g., maintenance manuals for automobile, in collaboration with a translation company in Japan. The result is promising. We could extract salient phrases which contain 80% of terms expected by the company. Currently, we try to get equivalents of extracted terms using SMT trained with the parallel documents.

5 Crowdsourcing Post-editing [6]

With properly controlled input sentences and substantial dictionary, state of the art MT system is useful, for example, for quick translations, such as news flash, and in-house translations. (Figure 1)

For document which needs higher quality, post-editing is required. However, post-editing can be costly and time-consuming, and not everybody can pay for it. We are conducting a preliminary investigation on the impact of crowdsourcing post-editing through the so-called "Collaborative Translation Framework" (CTF) developed by the Machine Translation team at Microsoft Research. Crowdsourcing translation is an increasingly popular-trend in the MT community, and we hope that our approach can shed new light on the research into crowdsourcing translation.

For our project, we used foreign students at Toyohashi University of Technology (TUT) and asked them to post-edit the MT output of TUT's websites (http://www.tut.ac.jp/english/introduction/) via Microsoft Translator into their own languages using the CTF functionalities. Though we do not expect that students have the same degree of accuracy from the professionals, we can note that they have a better understanding of the context, and so this kind of collaboration could improve and reduce the cost of the post-editing process.

We finished first experiment using 22 foreign students attending our university to post-edit the MT output of English version of the university's Web site into their own languages. Currently, we are conducting an experiment using 4 Japanese students with more precise settings, such as ordering of post-editing.

6 Concluding Remarks

In this paper, we provided a brief overview of current research activities on machine translation, all of which are aimed to make machine translation practically useable. Though our research is not completed, some of the result obtained so far are promising.

Acknowledgement. This work was funded by the Strategic Information and Communication R&D Promotion Programme of the Ministry of Internal Affairs and Communications, Japan.

References

1. Fuji, M., et al.: Evaluation Method for Determining Groups of Users Who Find MT Useful. In: Proceedings of the Machine Translation Summit VIII (2001)
2. Isahara, H., et al.: Development of a Japanese-Chinese machine translation system. In: Proceedings of MT Summit XI (2007)
3. Tatsumi, M., et al.: Building Translation Awareness in Occasional Authors: A User Case from Japan. In: Proceedings of EAMT 2012 (2012)
4. Hartley, A., et al.: Readability and Translatability Judgments for "Controlled Japanese". In: Proceedings of EAMT 2012 (2012)
5. Yamamoto, E., et al.: Extraction of Informative Expressions from Domain Specific Documents. In: Proceedings of LREC 2008 (2008)
6. Aikawa, T., et al.: The Impact of Crowdsourcing Post-editing with the CollaborativeTranslation Framework. In: Proceedings of JapTAL 2012 (2012)