Application of Resource-based Machine Translation to Real Business Scenes

Hitoshi ISAHARA[1,2], Masao UTIYAMA[1], Eiko YAMAMOTO[2,1],

Akira TERADA[3], Yasunori ABE[3]

[1]National Institute of Information and Communications Technology
3-5 Hikari-dai, Seika-cho, Souraku-gun, Kyoto 619-0289, Japan
[2]Graduate School of Engineering, Kobe University
1-1 Rokkodai-cho, Nada-ku, Kobe, Hyogo 657-8501, Japan
[3]Japan Airlines Co., Ltd.
Terminal 1, 3-2 Haneda Airport 3-chome, Ota-ku, Tokyo 144-0041, Japan
E-mail: isahara@nict.go.jp, mutiyama@nict.go.jp, eiko@mech.kobe-u.ac.jp, akira.terada@jal.com, yasunori.abe@jal.com

Abstract

As huge quantities of documents have become available, services using natural language processing technologies trained by huge corpora have emerged, such as information retrieval and information extraction. In this paper we verify the usefulness of resource-based, or corpus-based, translation in the aviation domain as a real business situation. This study is important from both a business perspective and an academic perspective. Intuitively, manuals for similar products, or manuals for different versions of the same product, are likely to resemble each other. Therefore, even with only a small training data, a corpus-based MT system can output useful translations. The corpus-based approach is powerful when the target is repetitive. Manuals for similar products, or manuals for different versions of the same product, are real-world documents that are repetitive. Our experiments on translation of manual documents are still in a beginning stage. However, the BLEU score from very small number of training sentences is already rather high. We believe corpus-based machine translation is a player full of promise in this kind of actual business scene.

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 translation, 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.

But is machine translation really applicable to such situations? One study examined for what types of people current machine translation (MT) systems are useful (Fuji et al., 2001), by simulating the retrieval of web pages in a language different from one's mother tongue. However, there has been little research to verify the usefulness of MT systems in specific business situations.

In this paper we verify the usefulness of resource-based, or corpus-based, translation in the aviation domain as a real-world business. This study is important from both a business perspective and an academic perspective.

2. Corpus-based MT

Generally speaking, corpus-based MT requires huge amounts of text data. When we treat pairs of languages whose linguistic features are rather different, the amount of parallel text required increases. Compared with rule-based approaches which use the concept of classification, e.g. abstraction, SMT normally considers a sentence as a string of characters and performs poorly when the grammatical features of the input language and the output language differ widely. In comparison, EBMT uses dictionary lookup and even shallow syntactic analysis, and so is considered to be stronger than SMT in such situations.

We are therefore developing a syntactically augmented EBMT system between Japanese and Chinese for scientific documents (Isahara et al., 2007). The use of syntax analysis means that the head of a clause is treated not as a word, i.e. character string, but as an instance of a class.

One of the problems in machine translation is selecting proper equivalents. Although researchers insist that SMT and EBMT can cope with this problem provided there is sufficient training data, it is still not known whether SMT or EBMT, which mainly use not semantic (class) information but surface strings, can choose proper equivalents of fundamental terms in particular.

Intuitively, manuals for similar products, or manuals for different versions of the same product, are likely to resemble each other. Therefore, even with only a small training data, a corpus-based MT system can output useful

translations. The corpus-based approach is powerful when the target is repetitive. Manuals for similar products, or manuals for different versions of the same product, are real-world documents that are repetitive.

If we consider translation in a business context, i.e. translation in a company and/or translation as a business, there is frequent repetition. Documents such as reports and announcements are often written with a similar style, vocabulary and contents, and are translated repeatedly. Other examples of this type are manuals, which are revised frequently, and there may be similar manuals for similar products, with some parts of the manuals being identical.

3. MT in the Real World

Airline companies need to prepare many manuals for operating different aircraft, and the manuals may be very similar in style and contents. The leading aircraft company, Boeing, sells several series of aircraft, such as the 737, 747, 767, 777 and new 787. These may be similar in operation, and so their manuals are also similar. However, airline companies have traditionally translated each manual from scratch[1]. Machine translation technology could be usefully tailored to this kind of companies.

In this study, we used a collection of aviation operation manuals in Japanese and English: Minimum Equipment List and Configuration Deviation List MEL/CDL manuals (MCM), and Aircraft Operating Manuals (AOM). For example, we translated the texts in a manual by using all text in other manuals as resources.

In the study, we used the manuals for the 737, 747 classic, 747-400 and 767 as the training corpus, and the manuals for the 777 as the test corpus. We translated the Japanese text into English. This study has several interesting aspects.

Firstly, the use of previously translated documents as the basis for making new translations is often done in the translation process. Therefore, the process conducted in this study simulates actual translation tasks.

Nowadays, although most researches on corpus-based machine translation use the BLEU score for evaluation, it has the problem that it is very hard to generate many reference translations in different styles for evaluation. In contrast, in our study the target translation already exists in the form of manuals in English, so we need no other reference translations.

Another criticism of the BLEU score is that it is calculated as the average of correct n-grams and there is no concrete relation between the BLEU score and usefulness in a real-world task. Again, this problem does not exist in our study. There are at least two ways to use the output of MT systems, i.e. read the MT output for information acquisition such as personal translation of web pages, and use the MT output as a preliminary translation for post-editing and publishing. Our situation is the latter, so one of the evaluation criteria is the cost of post-editing, e.g., the number of key strokes, or edit distance, for post-editing. As shown in the following section, the BLEU score in our experiments are more than 40. In this situation, the BLEU score reflects how much of the MT output can be used in the final translation and therefore is related to the cost of post-editing.

Other features of our study were as follows;

There are many identical or similar sentences, syntaxes and terms in manuals, making them suitable for a corpus-based MT system. Therefore, even a small amount of parallel text effectively covers linguistic phenomena. In our experiments, 7,000 sentence training data can generate translations with more than 40 BLEU score.

Generally speaking, corpus-based, or automatic, machine translation has the merit that the same equivalents are always selected in the same context. Especially, terms in manuals have few variants for a concept, so it is easy to acquire proper equivalents from corpora.

Normally it is very hard for MT systems to do free translation, i.e. not word-for-word translation, however, if we limit the target documents to specific manuals, the system can learn such free translation from corpora, i.e. phrase-for-phrase translation. In addition, translation of double negatives into positive expressions, tuning for domains, and acquisition of long words are possible.

4. Statistical Machine Translation (SMT) Experiments

To verify the above-mentioned issues in corpus-based MT, we attempted to translate aircraft manuals as a real-world example.

We preprocessed AOM and MCM manuals for the Boeing 737, 747-classic, 747-400, 767 and 777. We first made a parallel corpus by aligning Japanese and English manuals using Uchiyama and Isahara's method (Utiyama & Isahara, 2003; Utiyama & Isahara, 2007). Japanese sentences were segmented by a Japanese morphological analyzer, ChaSen, and English sentences were tokenized and lowercased.

We used the Moses SMT system (Koehn et al., 2007), which is a phrase-based SMT system, to train and test our dataset. We made a phrase table from the training data. Usually, the lengths of phrases are limited to 7, for example. However, we included all phrases regardless of their length, because sentences in a manual are often reused in other manuals. By retaining long phrases, we can successfully translate such recycle sentences. We tuned

[1] Airline companies whose official languages are other than English have to make manuals in their own languages on the equipments and procedures for each type of airplane in their fleet first, and have to translate them into English in order to use such manuals in abroad.

Moses using the development data and tested it using the test data.

In the experiments explained in this paper, we created training data from the Boeing 747-400 and 767 MCM manuals, which contained 6800 English sentences and 5200 Japanese sentences.

We used the first 500 sentences of Boeing 777 MCM manual for tuning parameters and translated the remaining 2300 Japanese sentences into English.

We translated Japanese sentences into English sentences. The case-insensitive BLEU score (Papineni et al., 2002) for the test data was 44.46. This score is very high given the small size of the training data size. Consequently, it is very promising to apply SMT to this task. We also reviewed the translations and concluded that the outputs of the SMT system were relatively good. Some example translations and their references are shown in the Appendix.

Our experiments on translation of manual documents are still in a beginning stage. However, the BLEU score from very small number of training sentences is already rather high, and at least we can add two more MCM manuals for training data. We believe corpus-based machine translation is a player full of promise in this kind of actual business scene.

5. Conclusion

Corpus-based machine translation, such as EBMT and SMT, looks very promising in this area. Even with a small training corpus, we can generate proper output. As we already have similar manuals of five different series of airplanes, i.e. 737, 747-classic, 747-400, 767 and 777, we will conduct translation experiments with combination of manuals and will compare the results to the current results, in order to evaluate the effect of the amount of training data in these specific domain and documents.

Corpus-based MT is attractive because we simply need to prepare parallel text as examples. We would like to verify whether SMT (or EBMT) is applicable even to small training data that is sufficiently repetitive.

We intend to apply the method described in this paper to a new plane for which there is no translation yet. We will try English-to-Japanese translation for this task. To apply this method in the real world, we need to develop post-editing tools by which post-editor can easily copy-and-paste usable parts of the output of machine translation systems. This might be different from the existing tools for translation aids such as translation memories. As for evaluation, we will consider a method of calculating the edit distance automatically.

6. References

Fuji, M. et al. (2001). Evaluation Method for Determining Groups of Users Who Find MT Useful. In *Proceedings of the Machine Translation Summit VIII.*

Isahara, H. et al. (2007). Development of a Japanese-Chinese Machine Translation System. In *proceedings of the Machine Translation Summit XI.*

Utiyama, M., Isahara, H. (2003). Reliable Measures for Aligning Japanese-English News Articles and Sentences. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL).*

Utiyama, M., Isahara, H. (2007). A Japanese-English Patent Parallel Corpus. In *proceedings of the Machine Translation Summit XI.*

Koehn, P. et al. (2007). Open Source Toolkit for Statistical Machine Translation. In *proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session.*

Papineni, K. et al. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL).*

APPENDIX

We will list some of interesting results here.
The first line is an input Japanese sentence. The second line is with word number for the sake of explanation. The third line is the output by machine translation system. The number shows the origin of the part of output English sentence. For example, "these provisos are not intended to prohibit |12-16|" means that this part is the translation of 12^{th} to 16^{th} words in Japanese input, i.e. "に(12) 立ち入っ(13) て(14) も(15) よい(16)". The fourth line is the sentence in the original English manual, i.e. reference translation.

==========

Not word-for-word translation.

==========

Input: NOTE : 乗務 員 は 点検 の ため 当該 Crew Rest 内 に 立ち入っ て も よい 。

Input: NOTE(0) :(1) 乗務(2) 員(3) は(4) 点検(5) の(6) ため(7) 当該(8) Crew(9) Rest(10) 内(11) に(12) 立ち入っ(13) て(14) も(15) よい(16) 。(17)

Output: note : |0-4| these provisos are not intended to prohibit |12-16| in the |11-11| affected crew rest |8-10| inspections by crewmembers |5-7| . |17-17|

Reference: NOTE : These provisions are not intended to prohibit attendant rest inspections by crewmembers .

==========

Whole input sentence is translated into English directly.

==========

Input: 本 Kit は 、必要 により 航空機 部品 、工具 及び Material を 運ぶ 為 の Kit で ある 。

Input: 本(0) Kit(1) は(2) 、(3) 必要(4) により(5) 航空機(6) 部品(7) 、(8) 工具(9) 及び(10) Material(11) を(12) 運ぶ(13) 為(14) の(15) Kit(16) で(17) ある(18) 。(19)

Output: the purpose of this kit is to carry aircraft certified spare parts , tools and material as required . |0-19|

Reference: The purpose of this kit is to carry aircraft certified spare parts , tools and material as required .

==========

Compared with previous example, input sentence is divided into parts and is translated properly.

========== ====

Input: a) 対応 する Second Stage Turbin Bypass Valve を Open 位置 に 固定 する こと 。

Input: a(0))(1) 対応(2) する(3) Second(4) Stage(5) Turbin(6) Bypass(7) Valve(8) を(9) Open(10) 位置(11) に(12) 固定(13) する(14) こと(15) 。(16)

Output: a) associated |0-3| Second |4-4| stage |5-5| Turbin |6-6| bypass valve |7-8| is secured open , |9-16|

Reference: a) Associated second stage turbine bypass valve is locked open ,

==========

Translation of parallel structure.

==========

Input: この 場合 、EICAS Status Message COND TEMP SENSOR L または COND TEMP SENSOR R が 表示 される 。

Input: この(0) 場合(1) 、(2) EICAS(3) Status(4) Message(5) COND(6) TEMP(7) SENSOR(8) L(9) または(10) COND(11) TEMP(12) SENSOR(13) R(14) が(15) 表示(16) さ(17) れる(18) 。(19)

Output: in this case , |0-2| eicas status message |3-5| cond |6-6| temp |7-7| SENSOR |8-8| l or |9-10| cond |11-11| temp |12-12| SENSOR |13-13| r |14-14| will be displayed . |15-19|

Reference: In this case , EICAS Status Message COND TEMP SENSOR L or R will be displayed .