4-2 Efficient Technologies for Creating Parallel Corpora

UTIYAMA Masao

A large scale parallel corpus is essential language resource for corpus-based machine translation. However, it is very difficult to construct them. This paper discusses two methods for creating parallel corpora. The first is extracting parallel sentences from existing Japanese-English parallel documents. The second is supporting volunteer translators for creating new translations in order to obtain new parallel sentences, instead of gathering existing translations. These methods enable us to develop comprehensive parallel corpora.

Keywords
Machine translation, Language resources, Parallel corpora, Volunteer translator, Minna no Hon'yaku

1 Introduction

Large-scale parallel corpora are necessary language resources for corpus-based machine translation. It is however extremely difficult to create the corpora. In this paper we introduce two methods of creating large-scale parallel corpora. These methods can be applied to any languages but below we explain their application to Japanese-English parallel corpora.

The first method is to extract translated sentences from existing Japanese-English translation texts. The second is used not to look up the existing translation texts but to directly obtain new bilingual translations from supporting volunteer translators.

2 Extraction of translated sentences from existing Japanese-English translation texts

An example of Japanese-English translation texts are newspaper articles written in Japanese and translated into English. In many cases, however, the translated articles are not direct translations of the original articles. For example, the title of an English newspaper article, although actually written in English, is not a simple translation of the corresponding article written in Japanese but usually contains some modifications.

Another example of translation texts that contain a lot of noise is patents that are simultaneously applied for in Japan and the United States. These are called patent families. Since there are a great number of patent families, we can obtain large-scale parallel corpora from these families. The parallel corpora thus obtained can be used also to create a machine translation engine for patent translation. In fact, the parallel corpora made with the present method was used in the patent translation task of NTCIR-7 to 10 [1][2].

2.1 Patent family used to create parallel corpora

In this paper we construct a parallel corpora using patent data provided for the NTCIR-6 patent retrieval task [3]. The data consists of

- About 3.5 million Japanese patent applications published from 1993 to 2002, and
- About 1 million USPTO patents published from 1993 to 2000.

We found 85,677 patent families accord-

ing to the priority numbers assigned to the US patents. We examined the patent families and found that the sections "Detailed Description of Invention" and "Background of the Invention" are literal translations of each other in many cases. We thus decided to use these sections to construct our parallel corpora.

A simple pattern matching program was applied to the extracted patent families and a total of 149,603 sections of "Detailed description of invention" and "Background of invention" were obtained. In what follows, we call these sections "documents".

2.2 Alignment of sentences

In this study we used Utiyama and Isahara's method [4] for the alignment of bilingual sentences. The procedure is as follows. We first aligned the sentences of each document using standard dynamic programming [5]. In this operation we utilized the following similarity score for each document pair $J(i)$ and $E(i)$ [4].

$$\mathrm{SIM}(J(i), E(i)) = \frac{2 \times \sum_{j \in J(i)} \sum_{e \in E(i)} \frac{\delta(j,e)}{\deg(j)\,\deg(e)}}{|J(i)| + |E(i)|}$$

Here, j and e represent Japanese and English word tokens respectively. $|J(i)|$ and $|E(i)|$ are the number of Japanese words and English words in the i-th alignment respectively. $\delta(j,e)$ is 1 when j and e can be a translation pair and 0 otherwise. deg(j) represents the number of English words translated from j and deg(e) is the number of Japanese words translated from e. This similarity score is used to obtain a sentence alignment of optimal score by dynamic programming. Then the average similarity score of the Japanese document J and English document E is calculated by the following equation.

$$\mathrm{AVSIM}(J,E) = \frac{\sum_{i=1}^{m} SIM(J(i),E(i))}{m}$$

Here, $(J(1), E(1))$, $(J(2), E(2))$, ..., $(J(m), E(m))$ are the sentence alignments obtained by the dynamic programming. A high $\mathrm{AVSIM}(J, E)$ value occurs when the sentence alignments in J and E take high similarity values. In this case we consider that the documents J and E have a large degree of similarity.

The ratio of the number of sentences between the documents E and J was also used as its score.

$$\mathrm{R}(J,E) = \min\left(\frac{|J|}{|E|}, \frac{|E|}{|J|}\right)$$

Here $|J|$ and $|E|$ are the number of sentences in the document J and that in the document E, respectively. This score is high when the two documents contain similar numbers of sentences.

Finally, we defined the score of the documents $J(i)$ and $E(i)$ as

$$\begin{aligned}\mathrm{Score}(J(i), E(i)) = {} & \mathrm{SIM}(J(i), E(i)) \\ & \times \mathrm{AVSIM}(J,E) \\ & \times \mathrm{R}(J,E)\end{aligned} \quad (1)$$

This score is high when the degree of similarity of the sentences and that of the documents are both high.

2.3 Extraction of well-matched sentence alignments

After the application of the above method to 149,603 bilingual documents extracted in Subsection **2.1**, about 7 million sentence alignments were obtained. From these alignments, we extracted about 4.2 million one-to-one sentence alignments. We removed alignments whose Japanese sentences did not end with periods to obtain alignments of appropriate sentences. The resulting number of the obtained sentence alignments was about 3.9 million.

We sorted these alignments in decreasing order of scores and examined 20 sentence alignments ranked around 2,000,000 from the top. We found that 17 of the 20 alignments were almost literal translations of each other. We also examined 20 sentence alignments ranked around 2,500,000 and found that 13 of the 20 alignments were almost literal translations of each other. Based on these observations, we decided to extract the top 2 million sentence alignments to use as parallel corpora. Finally, we removed some sentences that contained more than 100 words and the sentences whose lengths were too imbalanced between Japanese and English. The number of sentence

alignments thus obtained was 1,988,732.

To check the validity of the translation of the extracted sentences, we extracted 1,000 sentence alignments randomly and examined the translation accuracy by the following two-step procedure. In the first step, we manually marked a sentence alignment as "A" if the Japanese and English sentences matched as a whole, "B" if these sentences had more than 50% overlap in their content, and "C" otherwise to evaluate the translation. The number of alignments marked as A was 973, B was 24, and C was 3.

In the second step, we manually marked an alignment as "A" if the English sentence reflected almost perfectly the content of the Japanese sentence, "B" if about 80% of the content was shared, "C" if less than 80% of the content was shared, and "X" otherwise. The number of alignments marked as A was 899, B was 72, C was 26, and X was 3. From these evaluations, we concluded that the sentence alignments in the parallel corpora were highly accurate translations.

Next we used these 1,000 sentence alignments to investigate the relationship between the human judgments and scores given in Equation (1), which are shown in Fig. 1. The figure shows the cumulative numbers of B, C and X against the ranks of sentence alignments sorted in decreasing order of the score. The solid line indicates that noisy alignments tend to have low ranks. Note that if noisy alignments are spread uniformly among the ranks, then the cumulative number of noisy alignments would follow the dotted line. Based on the results shown in the figure, we concluded that the score of Equation (1) could give a high score to appropriate sentence alignments.

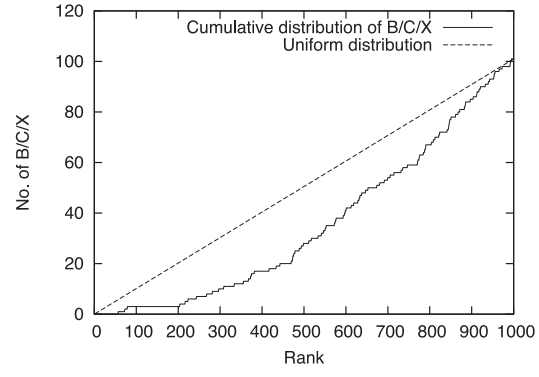

Fig.1 *Number of sentence alignments of B, C and X as function of score rank*

2.4 Machine translation experiments

Next we extracted 3.5 million sentence alignments from the above 3.9 million alignments by removing noisy alignments. We used the remaining alignments to study the relationship between the number of the sentence alignments used for the training of machine translation and the translation accuracy. For the machine translation experiment, we used Pharaoh decoder [6]. The translation accuracy was measured by %BLEU [7].

In order to evaluate the machine translation using %BLEU only by changing the number of sentence alignments used for the training, we used the same conditions in the experiment for the following: (1) Word alignments in translation data, (2) language models, (3) weight of translation model or language model, and (4) test data. We conducted the experiment only by changing the volume of the translation data of (1).

The common settings were obtained as follows. (1) The word alignments were calculated using GIZA++. (2) Trigram language models were made from the 3.5 million sentence alignments. (3) We commonly used, for all the settings, the feature weights that were obtained by training the models with the translation data of 2 million sentence alignments and training them with development data. (4) We commonly used 2,000 sentences as test data for all the settings.

Under these conditions, we calculated %BLEU by changing the training data volume from 0.5 million to 1 million, 1.5 million, 2 million, 2.5 million, 3 million, and 3.5 million sentence alignments. The results are shown in Fig. 2.

The figure shows that the %BLEU scores for the English-Japanese machine translation experiments reached a plateau around 2.5 mil-

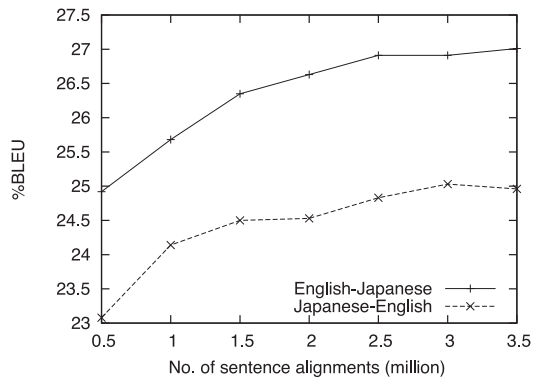

Fig.2 *%BLEU as function of number of sentence alignments used for training*

Fig.3 *Web site "Minna no Hon'yaku" (http://trans-aid.jp)*

lion sentence alignments. The %BLEU scores for the Japanese-English experiments increased for up to 3 million alignments and then dropped when 3.5 million alignments were used as the training data.

These observations indicate that, up to certain points, the increase in the size of the training data improves the accuracy of both English-Japanese and Japanese-English machine translations. However, the accuracy of the machine translations reached a plateau or even decreased after those points. Therefore we conclude that Equation (1) gives a high score to appropriate sentence alignments.

3 Acquisition of parallel corpora by supporting translation

The method introduced in Section **2** is used to create parallel corpora from existing bilingual texts. In this section, on the other hand, we introduce a method of obtaining parallel corpora from supporting volunteer translators.

Various translation support tools, such as the translation support editor QRedit [9], are available for supporting volunteer translators. However these tools are intended to be used for the support of individual translators, not for the simultaneous support of many translators.

The NICT MASTAR Project Multilingual Translation Laboratory developed a system for the simultaneous support of many translators over the world, in collaboration with the Library & Information Laboratory, University of Tokyo. In this project we created a Web site (http://trans-aid.jp) aiming to host volunteer translators. Figure 3 is a screen shot of our developed site "Minna no Hon'yaku" (Translation for All).

The reasons that we considered the hosting would be able to support many translators over the world are as follows.

(1) There are similar successful cases in other fields. For example, in the field of open sourcing, the development or distribution of open sources has been promoted by hosting open source projects, e.g. in sourceforge.net. We therefore expected that hosting volunteer translators would be successful.

(2) As seen from the reference [9], online volunteer translators do not often use translation support tools. The translation support site "Minna no Hon'yaku" has QRedit and other tools and hence provides users with an environment for using translation support tools spontaneously. The users are thus naturally supported in their translation works.

(3) In the site "Minna no Hon'yaku," the texts translated by volunteer translators are saved together with their original texts, so

that users of the site can not only see their own translated texts but also share texts translated by other translators. The translators can therefore use other's translations for their own translation work.

(4) In the site "Minna no Hon'yaku," the texts translated by a volunteer translator are saved and released to the public. The site therefore provides an opportunity to translators who want to release their translations.

For the above reasons, we considered that "Minna no Hon'yaku" could support many volunteer translators since, if they used the site, they could use the translation support tools and share and reuse translated texts.

"Minna no Hon'yaku" has the following characteristics: (1) Translators can use the highly-functional translation support editor QRedit. (2) Translated texts released in the site are licensed for the creation and release of derivative works from the texts. Therefore one can use the translated texts for appropriate purposes. (3) In cooperation with the Sanseido, "Grand Concise English Japanese Dictionary", 360,000 entries can be used for translation support.

3.1 Highly-functional translation support editor, QRedit

The basic design policies of the translation support editor QRedit are the following four points: (1) The editor does not provide new information or function but supports translators in their usual work. (2) It provides information necessary not for the system but for the translators to make a decision. (3) It shows information that would expand a translators' ideas. (4) It makes the translators' translation as simple as possible. These policies were determined on the basis of the interview to translators and the current level of the translation support technology.

With QRedit, translators can look up a word in multiple dictionaries and glossaries that they registered just by clicking the word to find a corresponding translation in a simple manner. It also has an advanced search function for idioms and can detect various idioms with a different word inserted.

QRedit shows idioms with underlines, as in Fig. 4, for translators not to overlook them. Since even skilled translators could translate an idiom incorrectly, this sort of warning from the editor is useful for them.

QRedit also has a Web search function and a term registration function. In particular, if terms are registered in QRedit, they can be looked up from the editor. One can look up not only the terms that he/she has registered but also those registered by others. Therefore if users of "Minna no Hon'yaku" register more and more terms and make them publicly available, one can find words that are not in ordinary dictionaries.

3.2 Sharing of translations

For the sharing of translations, it is necessary to consider the licensing of the original and translated sentences. For example, if a person who wrote a sentence does not allow others to make the translated sentence public, the translated sentence cannot be made public or shared, as a matter of course.

Users of "Minna no Hon'yaku" are hence requested to check the licensing of original and translated texts. They are also asked to allow secondary use of their translated texts. The system checks the licensing in the following manner when the users save their translated texts (see Fig. 5.).

(1) First, the system asks, "The text that you translated (original text) can be utilized only for private use or applications allowed by the Copyright Act, unless the author of the original text explicitly gives a permission. Does the author of the original text give you or others permission to make the translation of the text public?" (2) If the answer is Yes, the system asks translators to set a license condition that does not contradict the condition "Derivative works can be created from your translated text and made public," which is the condition of the Creative Commons Licenses.

"Minna no Hon'yaku" thus has a translation sharing system with emphasis on the

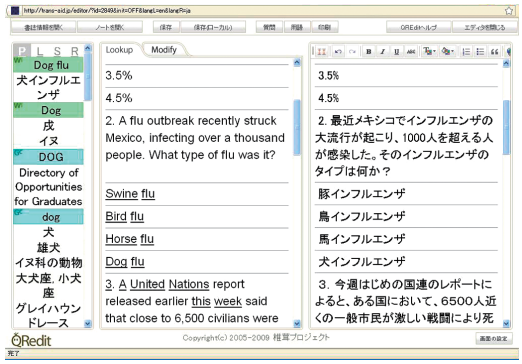

Fig.4 *Translation support editor QRedit*

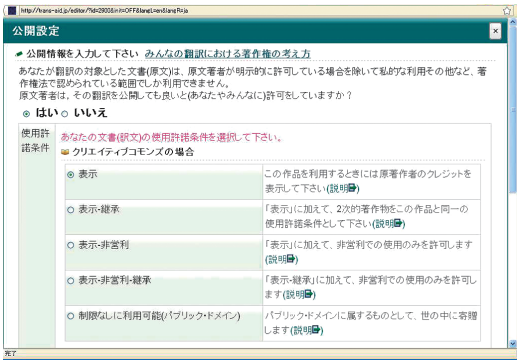

Fig.5 *Setting licensing of document*

copyrights of original text authors and translators.

3.3 Future development

"Minna no Hon'yaku" was released on April 8, 2009. The number of users as of June 2012 was more than 2100 and the number of translated texts was about 10,000. It also supports translation among the three languages of Japanese, English and Chinese, between English and Catalan, between Japanese and German, and between Japanese and Korean. In the future we would like to focus on educational uses of the site "Minna no Hon'yaku" for translators to become familiar with the site at an early stage.

4 Conclusions

In this paper we described two methods for creating bilingual texts. One is to extract bilingual texts from existing translation texts and the other is to obtain new bilingual texts by supporting translators. We can create comprehensive parallel corpora by the combination of these methods.

References

1 Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro, "Overview of the Patent Translation Task at the NTCIR-7 Workshop," Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, pp. 389–400, Dec. 2008.

2 Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro, Terumasa Ehara, Hiroshi Echizen-ya, and Sayori Shimohata, "Overview of the Patent Translation Task at the NTCIR-8 Workshop," NTCIR-8, pp. 371–376, 2010.

3 Atsushi Fujii, Makoto Iwayama, and Noriko Kando, "Overview of the Patent Retrieval Task at the NTCIR-6 Workshop," Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, pp. 359–365, May 2007.

4 Masao Utiyama and Hitoshi Isahara, "Reliable measures for aligning Japanese-English news articles and sentences," Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics, 2003.

5 William A. Gale and Kenneth W. Church, “A program for aligning sentences in bilingual corpora,” Computational Linguistics, 19(1): 75–102, 1993.
6 Philipp Koehn, “Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models,” Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA), pp. 115–124, 2004.
7 Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, “BLEU: A Method for Automatic Evaluation of Machine Translation,” Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL), 311–318, 2002.
8 Franz Josef Och and Hermann Ney, “A Systematic Comparison of Various Statistical Alignment Models,” Computational Linguistics, Vol. 29, No. 1, pp. 19–51, 2003.
9 Abekawa Takeshi and Kageura Kyo, “QRedit: An integrated editor system to support online volunteer translators,” Digital humanities, pp. 3–5, 2007.

Accepted June 14, 2012

UTIYAMA Masao, *Ph.D.*
Senior Researcher, Multilingual Translation Laboratory, Universal Communication Research Institute
Natural Language Processing, Machine Translation