Building a Biomedical Chinese-English Parallel Corpus from MEDLINE

Lingyi Tang, Jun Xu, Xinyue Hu, Qiang Wei, Hua Xu
School of Biomedical Informatics, The University of Texas Health Science Center at Houston
7000 Fannin St, Suite 870, Houston, TX U.S.A 77030
{Lingyi.Tang, Jun.Xu, Xinyue.Hu, Qiang.Wei, Hua.Xu}@uth.tmc.edu

Abstract

In this paper, we present a biomedical Chinese-English parallel corpus aligned at sentence level. We collected biomedical publications that are available in both Chinese and English from MEDLINE and generated a dataset of 5,129 bilingual abstracts. We then employed the Champollion aligner, which uses both lexicon and sentence length features, to align the sentences in both languages. The aligned parallel corpus contains 61,874 English sentences and 43,866 Chinese sentences. The corpus is still under development, as we are manually checking sentence boundary and the alignment. We believe such a publically available corpus will benefit the development of cross-lingual systems and applications in the biomedical domain.

Keywords: Parallel Corpus, Biomedical, Sentence Alignment, Chinese-English

1. Introduction

A parallel corpus is a collection of texts with the translation of one or more languages besides the original one, where the texts, paragraphs, sentences and words are typically linked to each other. The most common case of parallel corpora is bitext where only two languages are studied. Parallel corpora play a vital role in many natural language processing (NLP) tasks and applications, especially in Statistical Machine Translation (SMT). In addition, it is also a valuable resource for a wide range of multi-lingual research, such as cross-language information retrieval and information extraction. Parallel corpora can also serve as a reference to check whether the element of the words or phrases is correct when dealing with bi-lingual texts for translators and researchers.

Over the past several years, parallel corpora have been constructed widely in general domain, including resources for translation between English and Chinese texts. For example, the Institute of Computational Linguistics of Peking University has developed a large-scale Chinese-English Contrastive Language Knowledge Base （CECLKB）, containing 7.5 million words/characters (Bai et al., 2002). CECLKB is a sentence-aligned parallel corpus developed for formal description of Chinese and English sub-sentential comparison. It is in the XML format with contrastive knowledge for future implementation of NLP systems in a multilingual environment. Later, Bai et al. (2014) refined CECLKB with updated architecture, improved entry selection and implementation of XML-based annotation schemes. It serves as a general knowledge base for many NLP tasks such as Computer-Assisted Translation and Second Language Acquisition. In 2010, Mohammadi & GhasemAghaee (2010) proposed a Bilingual Parallel Corpora Base using Wikipedia. Researchers show increasing interest in Wikipedia because it is cooperatively edited, which makes its content up-to-date. In addition, it is entirely free and available in most of the languages in the world. Another advantage of leveraging Wikipedia for parallel corpus development is that its structures (e.g., definition sections followed by description sections) may remain similar to the same topic written in different languages. All these attributes have made Wikipedia a great resource for multilingual parallel corpora establishment. Another sizeable parallel corpus worth notice is the Hong Lou Meng Parallel Corpus developed by Yanshan University sponsored by National Social Science Foundation of China (Liu et al., 2008). It contains the original texts of 120 Chapters in Chinese and three representative translation texts in English which can be used for both independent or collaborative research. The whole corpus is aligned at sentenced level so that it can benefit cross-lingual information extraction and information retrieval. Another critical Chinese-English corpus is the Bilingual Corpora of Tourism Texts Corpus developed at the Hong Kong Polytechnic University (Li et al., 2010). Later Sun et al. (2014) further improved the bilingual tourism texts with travel guides, tourist information and travelogues. They demonstrated the use of thematic and formal features when translating Chinese tourism discourse text to English. However, as discussed in their paper, the translation in the tourism domain could be a relatively easy task, as the average word count of the original Chinese texts is usually similar to that in the translated English texts.

However, there is very limited work regarding building parallel Chinese-English corpora in the biomedical domain. One significant Chinese-English parallel corpus in the medical domain is PCMW (Chen & Ge, 2011), which contains texts from 15 English medical books, with corresponding translation texts in Chinese. Apparently, it is an excellent resource for developing medical Chinese-English machine translation systems as these books have a broad coverage of medical sub-domains. However, it is not open to public at this time. Nevertheless, there are multilingual biomedical corpora available in other languages, and they have been used for different NLP tasks. For example, Deleger et al. developed a word alignment method to automatically acquire translation between medical terms in English and French using existing parallel corpora (Deléger, Merkel, & Zweigenbaum, 2009). It would be interesting to conduct similar research to expand medical terminologies in

Chinese, if such biomedical Chinese-English parallel corpora are available.

This paper describes our effort on building a biomedical Chinese-English parallel corpus using MEDLINE abstracts that are available in both Chinese and English. We applied and evaluated two different algorithms to automatically construct the aligned parallel corpus at sentence level.

2. Method

This study attempts to automatically build a parallel corpus using available bilingual abstracts from MEDLINE. We downloaded and extracted all available bilingual abstracts from PubMed and then applied two different sentence alignment approaches to align sentences. To evaluate the performance of sentence alignment, we annotated a small dataset and used it to report the performance of each method. We plan to continue manually reviewing sentence alignments, thus to improve the quality of the parallel corpus.

2.1 Data Collection

We used the query "Chinese[Language]" to search biomedical abstracts that are originally in Chinese from the PubMed. Then we downloaded all MEDLINE entries (including titles and abstracts) of relevant publications in the search results using the Entrez Programming Utilities (E-utilities) [1]. We obtained a collection consisting of 289,597 publications in XML format. We then extracted Chinese and English abstract pairs by parsing the XML documents. Most of the records contain abstracts in a single language only. Finally, we obtained a dataset containing 5,129 pairs of abstracts in both English and Chinese.

2.2 Preprocessing

After a bilingual abstract was extracted from XML document, further preprocesses were applied. For English abstracts, a regular expression-based sentence boundary detection program and a tokenization program developed in our lab were used to break each English abstract into sentences and tokens (Soysal et al., 2017).

Some Chinese abstracts in the collection were written in traditional Chinese. For consistency, we converted them to simplified Chinese (used in mainland China) using OpenCC[2], an open source simplified-traditional Chinese conversion tool. We split the abstracts into sentences using punctuation marks such as "。", "？" and "！"., which are used to indicate a full-stop of a sentence. Since Chinese is standardly written without spaces between words, we used JieBa [3] Word Segmenter to split a sentences into a sequence of words.

2.3 Sentence Alignment

Researchers have proposed a number of automatic sentence alignment approaches, mainly in two categories: length-based and lexical-based. Length-based algorithms are simple: it aligns sentences based on their length in terms of words or characters, such as the Gale-Church algorithm (Gale et al., 1993). The Gale-Church algorithm's assumption is that long sentences in original language should be long sentences in the translated texts, while short sentences tend to be short in the translated texts. The algorithm uses dynamic programming to search the best alignment using the probabilities produced for each bead based on the sentence length. It performs well on texts written in alphabetic languages.

However, length-based alignment algorithms do not perform well on aligning syllabic/logographic language (e.g., Chinese) to alphabet language (e.g., English)(Tan et al., 2014). The Champollion algorithm uses both sentence length features and lexicon features to align two sentences. It borrows the idea of *tf-idf* weight to calculate the similarity between two text segments. *tf-idf* is a statistic that reflects how important a term is to a document in a corpus, which has been widely used in information retrieval tasks. The Champollion algorithm also defines the segment-wide term frequency (*stf*), i.e. the number of occurrences of a term within the segment. Thus, the *stf* is able to measure the importance of the term within the particular segment. The algorithm uses the combination of *stf* and *tf-idf* to evaluate the importance of a translation term pair. A higher weight will be assigned to a less frequent translation term pair because these term pairs provide stronger evidence for aligning two segments. For instance, in the following sentence pair in one abstract:

Chinese: **出院** 时仅 有 5 例 （ **7.1‰** ） 诊断 慢阻肺 。

English: And only 5 patients (**7.1‰**) were diagnosed as COPD at **discharge**.

We can tell that the pair (7.1‰, 7.1‰) is a stronger piece of evidence indicating the two sentences should be aligned because they appear less frequently at the same time. The pair ("discharge" and "出院") appears much more often than "7.1‰" in bilingual biomedical abstracts. Thus, the translation pair (7.1‰, 7.1‰) has a much higher weight than the pair of (discharge, 出院). This is in concordance with the observation.

For any two segments, the Champollion algorithm calculates the similarity score based on the term pair weight, the number of sentences, and the sentence length. Moreover, the dynamic programming algorithm is used to search the path with maximum similarity, i.e., the prediction of the best alignments.

In this study, we used the Champollion algorithm as our primary sentence alignment method, as it has been reported with a high overall performance on English-Chinese sentence alignment task (Li et al., 2010). In addition, we also included the Gale-Church algorithm as a baseline.

2.4 Manual Correction

The automatic sentence alignment approach does not achieve a 100% accuracy. To build an accurate parallel corpus with sentences aligned, we implemented a manual review process on the top of the automated system. This process is still ongoing.

[1] https://www.ncbi.nlm.nih.gov/books/NBK25501/

[2] http://opencc.byvoid.com

[3] https://github.com/isuhao/jieba

2.5 Evaluation

To evaluate the sentence alignment algorithm, we constructed a gold standard dataset of 100 English-Chinese abstract pairs, which were randomly selected from the entire corpus and manually aligned. The review was performed by a native Chinese speaker who is also fluent in English. The results of sentence alignment were measured in Precision, Recall and F-measure. The formula for each measurement is listed as below.

$$\text{Precision} = \frac{\#\text{Correct\_links}}{\#\text{Predicted\_links}}$$

$$\text{Recall} = \frac{\#\text{Correct\_links}}{\#\text{Reference\_links}}$$

$$\text{F-measure} = 2\frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

3. Results and Discussion

Table 1 shows the descriptive statistics of the constructed parallel corpus. The generated corpus contains 5,129 bilingual abstracts collected from the biomedical domain. There are 61,874 English sentences and 43,866 Chinese sentences.

#Documents	5,129
# English sentences	61,874
# English tokens	1732K
# Chinese sentences	43,866
# Chinese tokens	1551K

Table 1 The statistics of the bilingual corpus

Type	Number	Percentage(%)
1→1	25,041	60.62
2→1	6,595	15.97
3→1	3,529	8.54
4→1	2,484	6.01
0→1	1,364	3.30
1→2	1,285	3.11
1→3	181	0.44
1→4	46	0.11
2→2	771	1.87
Others	13	0.03
Total	41,309	100

Table 2 The statistics of aligning English sentences to Chinese sentences in the corpus

Table 2 shows the numbers of different types of alignment between English and Chinese sentences. In this 'silver' standard corpus, we can find that almost 30% aligned sentences are the n→1 type. It is reasonable as Chinese texts tend to concatenate multiple clauses with commas. As a result, a corresponding Chinese sentence could consist of several English sentences in the original text.

Algorithm	Precision	Recall	F-measure
Champollion	0.8079	0.8338	0.8206
Gale-Church	0.2862	0.2954	0.2907

Table 3: The performance of the sentence alignment algorithms on the test dataset

Table 3 shows the results of the Champollion and Gale-Church algorithms on aligning sentences from English to Chinese using the test dataset of 100 manually reviewed abstracts. Our experiment results show that it is challenging for the Gale-Church algorithm to align syllabic/logographic language to alphabetic language, as it merely uses the statistic information from the unaligned text. The performance of the Champollion method is much better than the Gale-Church algorithm. However, the precision and recall are noticeably lower than those reported on the datasets from general domain, indicating it is more challenging to align sentences in biomedical texts. There could be several reasons for this finding, such as lack of Chinese-English medical term pairs, low performance of the Chinese segmentation program in biomedical text, mismatches of lengths of terms (e.g., abbreviations) etc. Further analysis is needed to accurately identify reasons for such errors and to identify potential solutions.

4. Conclusion

In this paper, we developed a parallel biomedical corpus with aligned sentences using a collection of bilingual abstracts collected from MEDLINE and an automated sentence alignment algorithm. The corpus contains 5,129 bilingual abstracts, 61,874 English sentences and 43,866 Chinese sentences. Our evaluation using a manually reviewed test set of 100 bilingual abstracts shows that the sentence alignment algorithm achieved an F-measure of 0.8206. We believe this parallel corpus will benefit the development of Chinese-English translation systems and applications in the biomedical domain.

The corpus is still under development: we are manually checking the sentence boundary and alignments generated by the automated algorithm. We will release the final corpus to the public once it is done. In the future, we will keep expanding it. About 80% abstracts in the original collection are written in English only. However, it is possible to further query Chinese bibliographic databases to find those abstracts in Chinese, thus expanding the parallel corpus.

5. References

Bai, X., Chang, B., Zhan, W., & Wu, Y. (2002). The construction of a large-scale Chinese-English parallel corpus. *In Recent Development in Machine Translation Studies-Proceedings of the National Conference on Machine Translation* (pp. 124-131).

Bai, X., Zähner, C., Zan, H., & Yu, S. (2014). The Chinese-English Contrastive Language Knowledge Base and Its Applications. *In Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data* (pp. 107–119). Springer, Cham. https://doi.org/10.1007/978-3-319-12277-9_10

Deléger, L., Merkel, M., & Zweigenbaum, P. (2009). Translating medical terminologies through word alignment in parallel text corpora. *Journal of Biomedical Informatics*, 42(4), 692-701.

Gale, W. A., & Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational linguistics*, *19*(1), 75-102.

Gurulingappa, H., Rajput, A. M., Roberts, A., Fluck, J., Hofmann-Apitius, M., & Toldo, L. (2012). Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of biomedical informatics*, *45*(5), 885-892.

Kim, J. D., Ohta, T., Tateisi, Y., & Tsujii, J. I. (2003). GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, *19*(suppl_1), i180-i182.

Li, D., & Wang, K. (2010). Development and application of bilingual corpora of tourism texts: A new approach [J]. *Modern Foreign Languages*, *1*, 007.

Li, P., Sun, M., & Xue, P. (2010, August). Fast-Champollion: a fast and robust sentence alignment algorithm. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 710-718). Association for Computational Linguistics.

Liu, Z., Tian, L., & Liu, C. (2008). The compilation of Hong Lou Meng Chinese-English parallel corpus [J]. *Contemporary Linguistics*, 4, 007.

Mohammadi, M., & GhasemAghaee, N. (2010). Building Bilingual Parallel Corpora Based on Wikipedia. *In 2010 Second International Conference on Computer Engineering and Applications* (Vol. 2, pp. 264–268).

Soysal, E., Wang, J., Jiang, M., Wu, Y., Pakhomov, S., Liu, H., & Xu, H. (n.d.). CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*. https://doi.org/10.1093/jamia/ocx132

Tan, L., & Bond, F. (2014). NTU-MC toolkit: Annotating a linguistically diverse corpus. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations* (pp. 86-89).

Sun, Y., & Tang, F. (2014). A Parallel Corpus-based Investigation of Vocabulary Features of Tourism Translations1. *International Journal*, 2(3), 01-22.