UTHealth
The University of Texas
Health Science Center at Houston
School of Biomedical Informatics

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 USA 77030
{Lingyi.Tang, Jun.Xu, Xinyue.Hu, Qiang.Wei, Hua.Xu}@uth.tmc.edu

Introduction

- **Parallel Corpus**
  - 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
  - Play a vital role in **natural language processing (NLP)**
  - Especially in **statistical machine translation (SMT)**
- Very limited work regarding building parallel Chinese-English corpora in the biomedical domain.

Goal

- Provide valuable resource for a wide range of multi-lingual research, such as cross-language information retrieval and information extraction.
- 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.

Data Collection

1. Use the query "Chinese[Language]" on PubMed
2. Downloaded all MEDLINE entries (including titles and abstracts)
3. 289,597 publications in XML format
4. Extracted Chinese and English abstract pairs by parsing the XML documents
5. 5,129 pairs of abstracts in both English and Chinese.

Preprocessing

English abstracts:

- Regular expression-based tokenization and sentence boundary detection (Soysal et al., 2017).

Chinese abstracts:

1. Converted to simplified Chinese (used in mainland China) using OpenCC
2. Split sentences using punctuation marks (e.g., “。”, “？” and “！”
3. JieBa Word Segmenter to split sentences into a sequence of words.

Sentence Alignment

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**.

- Baseline: Gale-Church algorithm
  - Dynamic programming search for best alignment using the probabilities produced for each bead based on the sentence length.
- Primary method: Champollion algorithm
  - Uses both sentence length features and lexicon features to align two sentences.

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.

Evaluation

- Constructed a gold standard dataset of 100 English-Chinese abstract pairs, randomly selected from the entire corpus and manually aligned

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

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

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

Statistics of the Bilingual Corpus

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

Statistics of Aligning English sentences to Chinese sentences

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

Performance of the Sentence Alignment Algorithms

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

Conclusion

- A parallel biomedical corpus with aligned sentences from 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.
- *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.

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.