ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 2, No. 1, pp. 229-235, January 2011

doi:10.4304/jltr.2.1.229-235

Length-based Bilingual Sentence Alignment in Different Literary Forms

Jun Zhao
School of Foreign Languages, Tianjin University of Technology, Tianjin, China
Email: junzhao.tj@gmail.com

***Abstract*—Text alignment can be a useful practical tool for assisting translators. We use length-based method to calculate and try to find out the estimated correlation between different literary forms. Also the findings can be revealed from the modern and classical Chinese texts. At last we get a potential lexical pattern from different texts following the basic lexical-based method.**

***Index Terms*—sentence alignment, length-based method, literary form, content-focused text, form-focused text, classical Chinese**

I. Introduction

The possible units of translation are phonemes, morphemes, words, phrases, sentences and entire texts. Chinese-English sentence alignment method taking word as length computation unit gets the best alignment result. The Chinese-English corpora that are translated literally are suitable to be aligned on sentence level with statistical method. The following experiments are carried out through POS tagging, frequency and rank detecting. We use length-based method to calculate and try to find out the estimated correlation between different literary forms. Also the findings can be revealed from the modern and classical Chinese texts.

II. Literature Review

Most of the available length-based Chinese-English sentence alignment methods take Byte as their sentence length computation unit. The Chinese-English corpora that are translated literally are suitable to be aligned on sentence level with statistical method. But the method used to compute the parameters in processing Indo-European language must be adjusted before applying to Chinese-English corpora. Zhang Xia, Zan Hongying and Zhang Enzhan (2009) proposed six different sentence length computation methods which take verb, noun, adjective, content word, byte and all the word of a sentence is proposed as the sentence length. The experiment results illustrate that Chinese-English sentence alignment method taking word as length computation unit gets the best alignment result, and the precision and recall are 99.01% and 99.5% respectively. Lv Xueqiang, Li Qingyin, Huang Zhidan, Shen Yanna and Yao Tianshun (2008) presented five evaluation functions based on the former two evaluation functions, and then seven functions are researched comparably. Lai Maosheng and Qu Peng (2009) analyzed the query logs from search engine. Pos tagging is used to get the characters of high frequency POS (part of speech) results. Web users' use nouns to do concept focused retrieval and keywords are still the primary method to search on the Web.

III. Theoretical Background

Translation is not only a science with its own laws and methods but also an art of reproduction and recreation. Good translation conforms to the principles—faithfulness in content, expressiveness in wording and closeness in style. The possible units of translation are phonemes, morphemes, words, phrases, sentences and entire texts. The possible units of translation are phonemes, morphemes, words, phrases, sentences and entire texts. Translating consists in reproducing in the receptor language the closest natural equivalent of the source language message, first in terms of meaning and secondly in terms of style. (Eugene A. Nida) A word is a world. It is history in the briefest form. It is a spot on a page but often a story of great events and movement. You can't examine a word and learn it well without learning more than a word. (Charles W. Ferguson)

Indeed it is difficult to align translations based on the words, for it is more difficult to decide which words in an original are responsible for a given one in a translation and some words apparently translate morphological or syntactic phenomena. But text alignment can help us dealing with EST translation and reduce the burden on human by first aligning the old and revised document to find out changes, then aligning the old document with its translation, and finally splitting in changed sections in the new document into the translation of the old document, in which way the translators only have to translate the changed sections. The commonest case of one sentence being translated as one sentence is referred to as a 1:1 sentence alignment. Studies suggest around 90% of alignment are usually of this sort. But sometimes translators break up or join sentences, yielding 1:2 or 2:1, and even 1:3 or 3:1 sentence alignments.

The following experiments are carried out through POS tagging, frequency and rank detecting by ICTCLAS developed by Chinese Academy of Sciences and CLAWS (Constituent Likelihood Automatic Word-tagging System) developed by Lancaster University Centre for Computer Corpus Research on Language. CLAWS has consistently achieved 96-97% accuracy (the precise degree of accuracy varying according to the type of text). And the corpora are the latest (WEO) world economic outlook projection released at July 8, 2010 from World Bank, several chapters from *The Importance of Living* and *Quiet Dream Shadows by* Lin Yutang. Dr. Lin was very active in the popularization of classical Chinese literature in the West, as well as the general Chinese attitude towards life. His informal but polished style in both Chinese and English made him one of the most influential writers of his generation, and his compilations and translations of classic Chinese texts into English were bestsellers in the West. *The importance of living* published in 1937 was an outstanding translation work expressing the disposition culture to help westerners get better knowledge of Chinese culture. It was originally written in modern Chinese by Lin Yutang. *Quiet Dream Shadows* by Zhang Chao, a well-known writer from Qing Dynasty, was translated by Lin Yutang. The results show the similarity and disparity through the comparison between different literary forms, practical writings and literary works. Also the findings can be revealed from the modern and classical Chinese texts. There are about 150 pairs of sentences of each category in this experiment. We use length-based method to calculate and try to find out the estimated correlation between different literary forms. At last we get a potential lexical pattern from different texts following the basic lexical-based method.

IV. Methods and Findings

The principal kinds of text in the content-focused type would include press releases and comments, news report and other technical fields such as EST (English for Science & Technology or Technical English of Scientific English). It has its own stylistic features due to the specialty in content, field and discourse functions, and partly due to the unique habits of writers, which are mostly represented in lexical level and syntactical level being simply put as lengthy words (compounds, abbreviations, pseudo-technical terms and logic connectors), nominalization, the present and the perfect tense, the passive voice, the antecedent 'it' construction and double or triple propositions. We categorize the four groups, namely full words (content words), adjectives, verbs and nouns as different standards to identify the correlation between the English and Chinese texts. Adjectives are the least closely correlated tagging group next to verbs and nouns. The scatter plot can be perfectly described under the full words analysis.

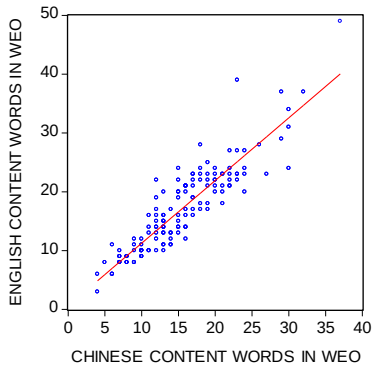

Figure 1. English Content Words in WEO vs. Chinese Content Words in WEO

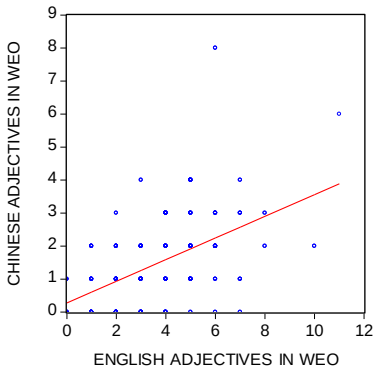

Figure 2. English Adjectives in WEO vs. Chinese Adjectives in WEO

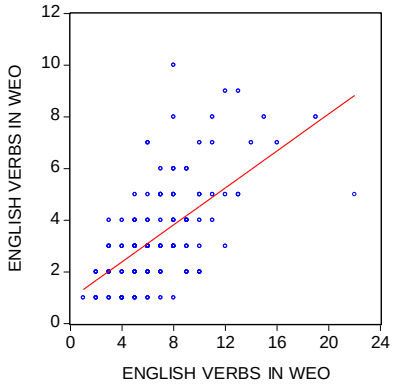

Figure 3. English Verbs in WEO vs. Chinese Verbs in WEO

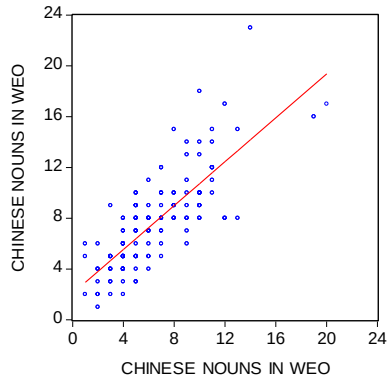

Figure 4. English Nouns in WEO vs. Chinese Nouns in WEO

Form-focused texts include literary prose, imaginative prose, and poetry in all its forms. Except for the adjectives which are the most irregular group, the verbs are now less highly correlated compared to the neatly plotted verbs WEO text. The first of all full words plot is less strongly correlated which is fitted in form-focused texts. As a transition part of the experiments, we mostly focus on the several literary forms and language revolutions results.

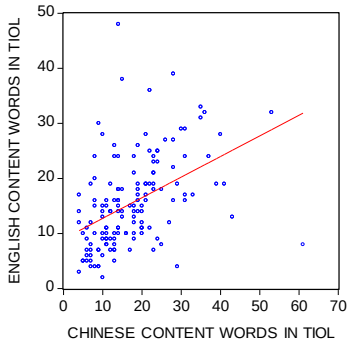

Figure 5. English Content Words in TIOL vs. Chinese Content Words in TIOL

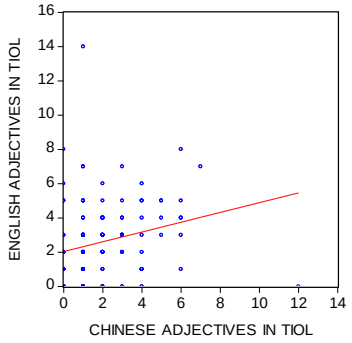

Figure 6. English Adjectives in TIOL vs. Chinese Adjectives in TIOL

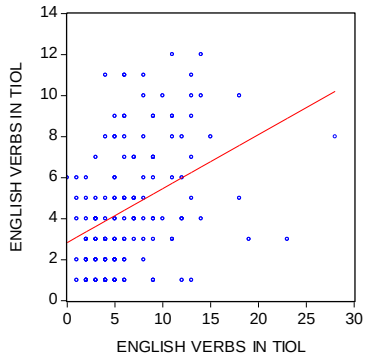

Figure 7. English Verbs in TIOL vs. Chinese Verbs in TIOL

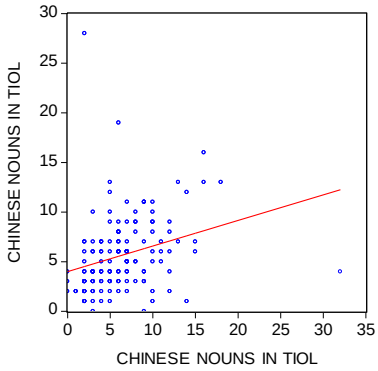

Figure 8. English Nouns in TIOL vs. Chinese Nouns in TIOL

Classical Chinese has no development of a prescriptive grammar and only by imitating earlier models rather than by obeying explicit rules as in Latin. Full words, also called content words and empty words are two major categories in Classical Chinese. The full words include nouns, verbs, adjectives, numerals and unit words, also known as expressions of quantity. On the contrary, the empty words relates to particles with its main function carrying grammatical meanings. The foreigners beginning their own research or learning often find themselves at a loss reading "real" texts, in which such problems abound. Much more important than any of the secondary literature, however, is to read the sources themselves, in translation at first and later in the original. With the exception of the subject and predicate part which both are the necessary parts of the sentences, in Chinese there are some special features among the sentences. Some of the verbs are used as noun-like words whereas other words are placed as preposition usage in English but called co-verbs. Numerals and expression of quantity also behave syntactically like verbs.

The last group in our analysis is the following estimation. It tells us that in this section verbs and nouns are sparsely scattered and the dots which are decentralized around the regression line increase. On the contrary, the full words plot seems much better applicable to the prediction than the form-focused texts. According to the diverse characteristics of the languages and authors, the Classical Chinese works had better be co-translated by both translators at home or abroad.Since it is also difficult when we translate Classical Chinese into modern Chinese, there are more considerations should be taken into account when we translate or interpret Classical Chinese into English. Similar to the Old English, lexical and syntactical traditions must be the two misleading factors through the translation. A more detailed classification could be set into the Classical literature as to the everyday usage and literary usage and then a more useful conclusion could be drawn.

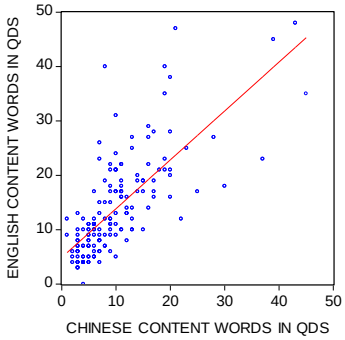

Figure 9. English Content Words in QDS vs. Chinese Content Words in QDS

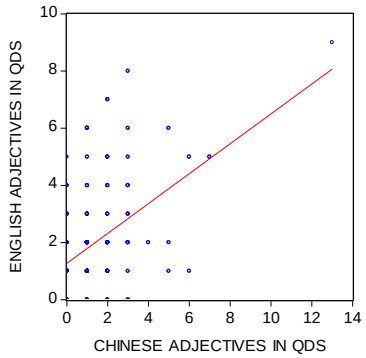

Figure 10. English Adjectives in QDS vs. Chinese Adjectives in QDS

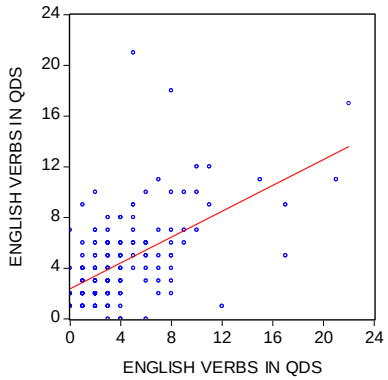

Figure 11. English Verbs in QDS vs. Chinese Verbs in QDS

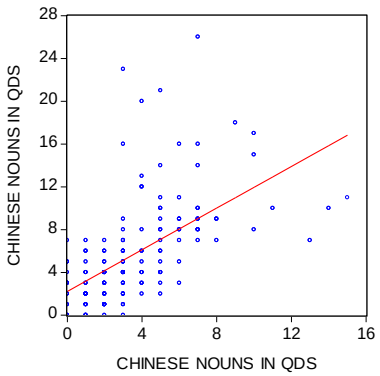

Figure 12. English Nouns in QDS vs. Chinese Nouns in QDS

Zipf's law states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table. It is not known why Zipf's law holds for most languages. However, it may be partially explained by the statistical analysis of randomly-generated texts. We only make the Classical Chinese rank-frequency analysis in order to get a specific result under our assumption. The distribution applies to the law under various empirical estimations. The graph shows rank on the X-axis versus frequency on the Y-axis. The Zipf's law graph corresponds to Zipf's prediction that the frequency of the most frequent words is lower. Here we only present the regular graph rather than the doubly logarithmic axes which approximates the second best result. In the future, a further analysis related to the Chinese poetry maybe conducted but only with the improvement of the tagging or other tools.

$$f \propto \frac{1}{r}, \tag{1}$$

or, in other words:

There is a constant k such that

$$f \cdot r = k .$$

$$f = (r+\rho)^{-B} \text{ or } \log f = \log P - B\log(r+\rho). \tag{2}$$

Here P, B and ρ are parameters of a text, that collectively measure the richness of the text's use of words.

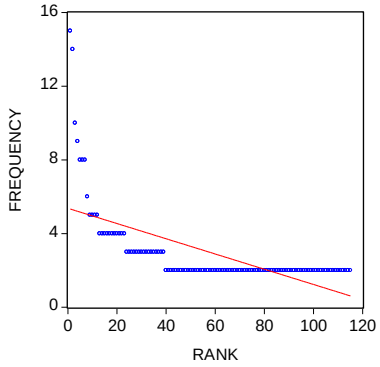

Figure 13. Frequency vs. rank

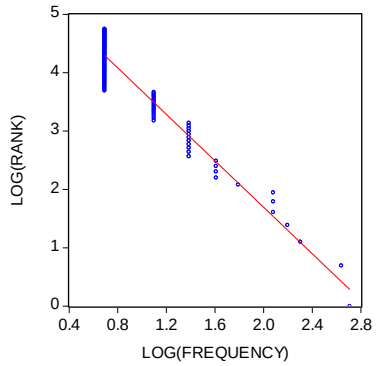

Figure 14. Log (rank) vs. log (frequency)

V. Conclusion

The results show the similarity and disparity through the comparison between different literary forms, practical writings and literary works. Also the findings can be revealed from the modern and Classical Chinese texts. Except for the adjectives which are the most irregular group, the verbs are now less highly correlated compared to the neatly plotted verbs WEO text. The first of all full words plot is less strongly correlated which is fitted in form-focused texts. Verbs and nouns are sparsely scattered and the dots which are decentralized around the regression line increase in Classical Chinese texts. On the contrary, the full words plot seems much better applicable to the prediction than the form-focused texts. Chinese-English sentence alignment method taking word as length computation unit gets the best alignment result. The Zipf's law graph corresponds to Zipf's prediction that the frequency of the most frequent words is lower.

Acknowledgment

The author wishes to thank Prof. Xu for his kindness and help.

References

[1] Catford, J. C. (ed.) (1965). A linguistic theory of Translation. Oxford: Oxford University Press.
[2] Chris, Manning & Hinrich Schütze (eds.) (1999). Foundations of statistical natural language processing. MA: MIT Press.
[3] Free CLAWS WWW trial service. (2010). http://ucrel.lancs.ac.uk/claws/trial.html (accessed 8/8/2010).
[4] Halliday, M. A. K. (ed.) (2006). Studies in Chinese language. New York: Continuum International Publishing.
[5] Lai, M. S. & P. Qu. (2009). The POS and mining study on search engine's query log. *Knowledge Organization and Knowledge Management* 177.4, 50–56.
[6] Lin, Y. T. (ed.) (2009). The importance of living. BJ: Foreign Language Teaching and Research Press.
[7] Lin, Y.T. (ed.). (2009). The importance of understanding. BJ: Foreign Language Teaching and Research Press.
[8] Lin Y. T. (ed.) (2010). The importance of living. BJ: Qun Yan Press.
[9] Lv X. Q., Q. Y. Li, Z. D. Huang, Y. N. Shen & T. S. Yao. (2004). Towards Chinese-English sentence alignment based on statistical method. *Mini-micro Systems* 25.6, 990-992.
[10] Kay, M. & M. Rscheisen (eds.) (1993). Text-translation alignment. *Computational Linguistics* 19.1, 121-142.
[11] Nida, A. Eugene (ed.) (2001). Language and Culture: Contexts in translation. SH: Shanghai Foreign Language Education Press.
[12] Nida, A. Eugene (ed.) (2003). The theory and practice of translation. MA: Brill Academic Publishers.
[13] Pulleyblank, E. G. (ed.) (1996). Outline of Classical Chinese grammar. CA: UBC Press.
[14] UCREL CLAWS7 Tagset. (2010). http://ucrel.lancs.ac.uk/claws7tags.html (accessed 10/8/2010).

[15] Wang, F. (ed.) (2008). Quiet dream shadows. BJ: Zhong Hua Book Company.
[16] Wang, J. Q. & K. F. Wang. (2008). Translation studies from the perspective of computational linguistics. *Journal of Foreign Languages* 31.3, 78-83.
[17] World economic outlook update. (2010). http://www.worldbank.org/ (accessed 20/8/2010).
[18] World economic outlook update. (2010). http://www.worldbank.org.cn/Chinese/ (accessed 20/8/2010).
[19] Yang, Y. & G. Ma. (eds.) (2008). A practical study on EST translation. CN: Xi'an Jiaotong University Press.
[20] Zhang, X., H. Y. Zan & E. Z. Zhang. (2009). Study on length computation method of Chinese-English sentence alignment. *Computer Engineering and Design* 18, 4356-4359.

Jun Zhao was born in Tianjin, China in 1983. She has been working on her M.A. degree in translation in Tianjin University of Technology, China in 2008.

She worked as a graduate assistant, assisting and collaborating in academic teaching or research activities in China. Her research interests include translation ecology and natural language processing.