Hsin-Yi Lien

Validating Terminologies and Phraseological Units Retrieved from Specialized Comparable Corpora in Lexical Semantics: An Interactive Method

Keywords: terminology; phraseological units; comparable corpora; interactive method; lexical semantics

The retrieve of terminology and phraseology from a monolingual corpus currently performed effectively by tools but extraction of keywords, terms, multi-word patterns, or collocations remain challenging, whether parallel or comparable corpora are utilized. Bilingual terminology extraction is generally conducted using either parallel corpora (Ndhlovu, 2016) or comparable corpora (Terryn, Hoste, & Lefever, 2020), and most studies identify keywords, collocations, and terms using computational methods (Štajner, & Mladenić, 2019). Recent works have demonstrated that comparable corpora can be used in cases where parallel corpora is unavailable. Comparable corpora are significant, for they provide examples of attested usage in native-speaking contexts (Giampieri, 2018). In addition to balanced comparable corpora, Morin and Hasem (2015) used unbalanced specialized comparable corpora to examine the quality of extracted bilingual terminology through a regression model which word co-occurrences in the context were observed. Their results show that the quality of retrieved lexicons is good by using unbalanced specialized comparable corpora. Thus, the usability of comparable corpora in cross-language information retrieval is applicable in extracting bilingual terminology. Lien (2018) compiled unbalanced Buddhist comparable corpora and generate the keyword lists and collocation lists by using n-gram function in *Sketch Engine* and expert evaluation.

Most of the studies on extraction of bilingual terminology or lexical phrases employed mainly computational methods, such as word embeddings words combined with a kernel approximation (Štajner, & Mladenić 2019), STACC (Azpeitia, Etchegoyhen, & Garcia 2018), and *Sketch Engine* (Lien, 2018); however, the quality of obtained terminology or phraseological units in lexical semantic level was not evaluated in previous studies. Accordingly, the present study intends to utilize an interactive method to evaluate suitability of those retrievals from specialized comparable corpora and to analyze the terms and phraseological units in lexical semantics. The specialized comparable corpora consisted of a Buddhist English Corpus and a Buddhist Chinese Corpus. The comparable corpora used in the present study were Buddhist English Corpus (BEC) (Lien 2017) and Buddhist Chinese Corpus (BCC) (Lien, 2018). The *BEC* and *BCC* were both compiled from books, essays, e-books, articles and reviews. A total of 22,677,744 tokens were obtained in *BEC*. The corpus included four sub-corpora regarding to history (4,582,771), origins (2,161,005), beliefs (11,104,917), and arts (4,829,051). The *BCC* contained 20,318,513 tokens which were obtained from publicly available texts.

The four sub-corpora were Buddhist history (7,764,086), origins (3,092,059), beliefs (6,687,199), and arts (2,775,169). The two corpora were unbalanced specialized comparable.

The text files were converted to plain text (.txt) for further analysis. The methods of extracting terminologies from comparable corpora in previous studies were inclined to employ statistical machine translation or computational analysis. However, it was apparently insufficient for ensuring semantic level of obtained terms (Lien, 2022; Tongpoon-Patanasorn, 2018). Accordingly, the present study applied an interactive method for cross validation of the quality of retrieved terms. The proposed method included filtering the terms with criteria, validating terms with different references sources (google search engine, English dictionaries, Chinese dictionaries, Pali dictionaries), implementing various statistical measures (absolute frequency, LL, OR) for ensuring the distinctness of obtained terms, and machine translation for comparison of terminology and phraseological units. Moreover, the distinct terminologies and phraseological units extracted from specialized Buddhist comparable corpora were examined in lexical semantics. The change of trend in Eastern and Western Buddhist literature were explored through comparing the extracted terms and phraseological units from the Buddhist comparable corpora. Mutual information (MI) were utilized to attain collocates of key clusters which had the highest keyness values in their semantic functions occurring in the *BCC*. The researchers indicated it was a more suitable span for verbs and their collocates in text is (0, +5) as it covers most of the high-frequency collocations (Bai & Zheng, 2004; Li & Guo, 2016). The n-gram function in Sketch Engine were used to generate collocation list. Therefore, in the present study, a collocation was defined as a single word co-occurring in the span of ± 5 words from the reference word, co-occurring at least five times in total across at least five different texts with a MI score of at least 3 and a t-score of at least 2.

After the computational analysis was done, the manual review was used and those phraseological units which are not specific will be removed. To collocate the phraseological units retrieved from two corpora, the phraseological units extracted from BEC were translated into Chinese by using different reference sources: google search engine, English dictionaries, Chinese dictionaries, Pali dictionaries, *Princeton Dictionary of Buddhism* (Buswell & Lopez, 2014), *Digital Dictionary of Buddhism* (Muller, 2015). Some specific obtained phraseological units which may be Hindi or Pali, such as “calm abiding” which was a collocation appearing in BEC. It is “shamatha” in Pali and Chinese translation is “stillness”. The occurrence of those specific phraseological units in two lists were compared and analyzed in cultural perspectives.

References

Azpeitia, A./Etchegoyhen, T./Garcia, E. M. (2018): Extracting parallel sentences from comparable corpora with STACC variants. Proceedings of the 11th Workshop on Building and Using Comparable Corpora at LREC 2018, pp. 48-52.

Buswell Jr., R. E./Lopez Jr., D. S. (2014): The Princeton Dictionary of Buddhism. Princeton & Oxford: Prince ton University Press.

Bai, M./Zheng, J. (2004): Study on ways of verb-verb collocation. Computer Engineering and Applications, 27, pp. 70-72.

Giampieri, P. (2018): Online parallel and comparable corpora for legal translation. Altre Modernita-Rivista Di Studi Letterari E Culturali, 20, pp. 237-252.

Li, S./Guo, S. (2016): Collocation analysis tools for Chinese collocation studies. Journal of Technology and Chinese Language Teaching, 7(1), pp. 56-77.

Lien, H. Y. (2017): The analysis of religious corpus. Proceedings of the International Journal of Arts and Sciences, 10(2), pp. 305-306.

Lien, H. Y. (2018): Mining comparable corpora for cross-language information retrieval. CALL your Data, pp. 223-228.

Lien, H. Y. (2022): Revisiting keyword analysis in a specialized corpus: Religious terminology extraction. Journal of Quantitative Linguistics, 29(3), pp. 269-282.

Morin, E./Hasem, A. (2015): Exploiting unbalanced specialized comparable corpora for bilingual lexicon extraction. Natural Language Engineering, 22(4), pp. 575-601.

Muller, A. C. (2015). Digital Dictionary of Buddhism. Retrieved from http://www.buddhism-dict.net/dicts-intro.html

Ndhlovu, K. (2016): Using ParaConc to extract bilingual terminology from parallel corpora: A case of English and Ndebele'. Literator, 37(2), pp. 1-12

Štajner, T./Mladenić, D. (2019): Cross-lingual document similarity estimation and dictionary generation with comparable corpora. Knowledge and Information Systems, 58, pp. 729-743.

Terryn, A. R., Hoste, V., & Lefever, E. (2020): In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora. Language Resources and Evaluation, 54(2), 385-418.

Tongpoon-Patanasorn, A. (2018): Developing a frequent technical words list for finance: A hybrid approach. English for Specific Purposes, 51, pp. 45-54.

Buswell Jr., R. E./Lopez Jr. D. S. (2014): The Princeton Dictionary of Buddhism. Princeton & Oxford: Princeton University Press.

Giampieri, P. (2018): Online parallel and comparable corpora for legal translations. Altre Modernita-Rivista Di Studi Letterari E Culturali, 20, 237-252.

Li, S., & Guo, S. (2016): Collocation analysis tools for Chinese collocation studies. Journal of Technology and Chinese Language Teaching, 7(1), 56-77.

Lien, H. Y. (2017): The analysis of religious corpus. Proceedings of the International Journal of Arts and Sciences, 10(2), 305-306.

Lien, H. Y. (2018): Mining comparable corpora for cross-language information retrieval. CALL your Data, 223-228.

Lien, H. Y. (2022): Revisiting keyword analysis in a specialized corpus: Religious terminology extraction. Journal of Quantitative Linguistics, 29(3), 269-282.

Morin, E., & Hasem, A. (2015): Exploiting unbalanced specialized comparable corpora for bilingual lexicon extraction. Natural Language Engineering, 22(4), 575-601.

Muller, A. C. (2015): Digital Dictionary of Buddhism. Retrieved from http://www.buddhism-dict.net/dicts-intro.html

Ndhlovu, K. (2016): Using ParaConc to extract bilingual terminology from parallel corpora: A case of English and Ndebele'. Literator, 37(2), pp. 1-12

Štajner, T./Mladenić, D. (2019): Cross-lingual document similarity estimation and dictionary generation with comparable corpora. Knowledge and Information Systems, 58, pp. 729-743.

Terryn, A. R./Hoste, V./Lefever, E. (2020): In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora. Language Resources and

Evaluation, 54(2), pp. 385-418.

Tongpoon-Patanasorn, A. (2018): Developing a frequent technical words list for finance: A hybrid approach. English for Specific Purposes, 51, pp. 45-54.

Contact information

Hsin-Yi Lien

Graduate School of Education, Ming Chuan University

maggielien61@gmail.com