Using Chinese-English Parallel Corpora for Compiling Bilingual Legal Glossaries

DOI 10.6342/NTU201603217

作者 李海琳

關鍵字 名詞組編取；平行語料庫；法律用語；兩岸法律用語對比分析；n-gram phrase extraction；parallel corpora；legal expressions；contrastive analysis in Cross-Strait terminology

出版品名稱 臺灣大學翻譯碩士學位學程 學位論文

卷期出版年月 碩士

學位類別 碩士

指導教授 高振明

內容摘要 英文

Bilingual glossaries enable translators to maintain accurate domain-specific translation and consistency. This study aims to extract bilingual pairs semi-automatically and provides a systematic specialized term compilation for translators to follow. First, two natural language processing tools, Anymalign and Pialign, were adopted to extract Chinese-English candidate pairs that had higher translation probabilities from Taiwan’s and Mainland China’s criminal parallel corpora. As the extraction tools are not perfect, this study focuses on improving the quality of the preliminary extracted results: qualified English noun (phrases) were identified first. English words/phrases were assigned parts-of-speech labels by Stanford POS Tagger automatically. Linguistic information was helpful in identifying and removing non-noun (phrase) patterns in order to locate qualified English noun phrases (NP). 1,852 English NPs with their Chinese pairs were found to be qualified (Taiwan), while 3,782 NPs were found to be valid (Mainland China). These results include noun phrases with longer word units. Based on these candidate results, 694 bilingual pairs were identified to be correctly aligned (Taiwan), while 852 correctly aligned bilingual pairs were identified (Mainland China). An American reference corpus was utilized to mark terms out by the indication of Keyness scores. The threshold was set at the critical value $\geq 3.84$ calculated by the log-likelihood ratio. 394 terms were found among 487 qualified noun (phrases) from Taiwan, while 418 terms were identified among 517 Mainland China’s result. Qualified noun (phrases) and terms were adopted to produce a comparable list of noun (phrases), showing the similarities and differences in Chinese legal expressions and English translation across the Straits. Improving the extraction of English NPs was confirmed to be much more effective than that of Chinese NPs because Chinese parts-of-speech can be made clearer only in sentences. In contrast, as suffixes of English words are more distinct, it is shown that English noun (phrase) extraction was more effective by having identified their suffixes according to the parts-of-speech. Based on the result, extraction of bilingual pairs was proven to be difficult because, first, no regular patterns could be identified in the two main Chinese NP groups: those NPs include/exclude ‘之’ (in Taiwan’s NP), or ‘的/or ‘之’ (Mainland China’s NP). Second, the English translation patterns corresponding to Chinese NPs were neither always predictable. Filtering terms was proven to be even more difficult. As the length of an NP unit increases, mixture of terms and common words within an NP will more likely appear. Legal connotations are not entirely the same in Taiwan’s and Mainland China’s legal systems. If the word units of certain phrases are not extracted completely, falsely similar or seemingly different bilingual pairs will be created. The methodology and findings presented in this study are recommended to be applicable to other language pairs and different domain-specific genres, which may facilitate improvements in the compilation of bilingual term glossaries and localization in translation.
