漢語學術詞表提取

摘要：自 1953 年至今，在英語學習領域有關於詞表提取的相關研究已經形成了相當成熟的體系，尤其是針對於學術語料庫的學術詞表提取方法已形成了兩個流派，並針對不同學科也產生了許多成果。但類似的研究在漢語領域卻顯得略有不足，例如在詞彙學習的過程中我們並不會產生專門學習學術詞彙的概念，以及現有的相關研究相比英文領域比較落後，研究成果較為少見。但事實上，學術詞彙作為學術領域中十分重要的一部分，每一位學術研究人員都應當進行一定程度的學習，因此需要一份新學術漢語詞表來填補這一空白。我們根據 2021-2022 年中文社會科學引文索引（CSSCI）收集了 581 種權威中文期刊收錄的論文元資訊，構建了一個上億字元的學術語料庫，並以使用頻次、跨文本分佈、詞語均勻分佈值三個參數作為特徵與篩選條件，並在先前學術界研究的基礎上對使用特徵進行了進一步的改進，希望能夠得到在學術語料中“更重要”的詞彙詞表。我們最後應用計算語言學方法提取出了一個基於 2019-2022 年全部期刊論文摘要資料提取得到的通用學術漢語詞表，以及基於法學、管理學、教育學、語言學四個學科領域的最早自 1954 年以來至 2022 年發表論文的摘要資料組成的子語料庫提取得到四個特殊用途學術漢語的學科專用詞表，並在這五個詞表的基礎上使用漢語停用詞表（stop words）對重複的詞彙進行了排除，得到了刪除停用詞和不刪除停用詞兩個版本的結果，最後從語料庫覆蓋率、詞彙排序與重複詞分析三個方向與現有通用詞表對比進行了驗證。我們認為新學術漢語詞表的創新點不僅僅是學術漢語領域學術語料庫規模的提高以及通用學術詞表和多

個學科專用詞表的提出，同時也體現在包含英文的國際學術詞彙研究領域對詞語均勻分佈值這一指標改進演算法的應用。我們相信新學術漢語詞表的建設在詞彙研究、語料庫語言學、英語教學等多個領域都能給予相當的幫助和參考。

關鍵字：語料庫；學術詞彙；詞表；特徵提取

Extraction of Chinese Academic Vocabulary

Abstract: Since 1953, the related research on word list development has formed a fairly mature system in the field of English study. Especially for academic word list development based on academic corpus, two schools of academic word list development method have been formed, and many achievements have been made for different disciplines. However, similar studies in the field of Chinese seem to be slightly inadequate. For example, in the process of vocabulary learning, we will not realize to learning academic vocabulary specially, and the existing related studies are relatively backward compared with the English field, with relatively few results. In fact, as an important part of the academic field, academic vocabulary should be studied to a certain degree by every academic researcher. Therefore, a new Chinese academic vocabulary is needed to fill this gap. Based on the Chinese Social Sciences Citation Index (CSSCI) from 2021 to 2022, we collected meta-information of 581 authoritative Chinese journals, constructed an academic corpus with hundreds of millions of characters, used Frequency, Range, and Dispersion as the characteristics and screening conditions, and further improved the usage characteristics on the basis of previous academic research. We hope to find "more important" words based on academic corpus. Finally, we used computational linguistics method to developed a Chinese general academic word list based on the abstract data of all journal papers from 2019 to 2022, and four discipline word list which is Law, Management, Pedagogy and Linguistics based on sub-corpus from 1954 to 2022. And we remove repetitive words by Chinese stop words on these five word-list. The results of deleting the stop words and not deleting the stop words are obtained. Then,

the results are compared with the existing general word lists in three directions: the coverage of the corpus, the sorting of words and the analysis of repetitive words. We think that the innovative of the new Chinese academic word list is not only the improvement of the size of the academic corpus as well as the development of the general academic word list and the word list for many disciplines, but also the application of the improved algorithm to the index of word distribution value in the research field of international academic vocabulary. We believe that the construction of the new Chinese academic word list can provide considerable help and reference in many fields such as vocabulary research, corpus linguistics and English teaching.

Key words: Corpus; Academic Word; Word List; Feature Extraction