洪嘉馡
國立台灣大學語言學研究所
論文題目: 詞義預測研究：以語料庫驅動的語言學研究方法
指導教授: 黃居仁 教授、安可思 教授

論文摘要

In this study, I proposed using corpus-driven distribution as the main method of prediction. I concentrated on individual semantic features to predict the senses of non-defined words by using corpora and tools, such as Chinese Gigaword Corpus, HowNet, Chinese Wordnet, and XianDai HanYu CiDian (Xian Han). Using these corpora, I determined the collocation clusters of the four target words--- chi1 “eat”, wan2 “play”, huan4 “change” and shao1 “burn” through character similarities and concepts similarities.

The four target words are all transitive verbs and they each have more than two senses. The collocation words of the four target words are very useful and play an important role in this sense prediction study. When conducting the character similarity clustering analysis, I employed identical morphemes of some of the collocation words in order to cluster them into the same cluster. Therefore, there are two main strategies of the corpus-based and computational approach used in this sense prediction study: (1) character similarity clustering analysis; and (2) concept similarity clustering analysis, which encompasses via HowNet (a) similarity between sememes, and (b) similarity between concepts. In this sense prediction study, I first predicted that different clusters can represent different senses, and I examined the accuracy rates of the four target words via the character similarity clustering analysis and the concept similarity clustering analysis of the corpus-based and computational approach. Then, I evaluated the four target words via sense divisions in Chinese Wordnet and in Xiandai Hanyu Cidian and was able to employ automatically computational programming to predict different senses for chi “eat”, wan2 “play”, huan4 “change”, and shao1 “burn”.

After the corpus-based and computational approach used in this sense prediction study, I demonstrated that I was able to use off-line tasks to test my participants’ intuition, which supports the theory that different clusters can represent different senses when using the corpus-based and computational approach. Therefore, in order to examine the related collocation words for the lexically ambiguous target words, I employed a multiple-choice task (Burton et al. 1991). In addition, because the stimuli were collected from the character similarity clustering analysis of the corpus-based and computational approach, I demonstrated the viability of this

approach by the results presented in this sense prediction study.

B. 具體貢獻

Concerning the contribution of this study, there are three important main points as follows:

(1) First contribution: I provided different research approaches. Not only did I employ the common morphemes of words in order to cluster them into the same cluster in the character similarity clustering analysis, but also I utilized the concepts of the words in order to cluster them into the same cluster via concept similarity clustering analysis. In addition, I found that when I set 20-times predicting clusters as my default target for the four target words, they indeed followed reasonable distributions and presented the best results. Last but not least, I can predict physical senses and metaphorical senses of the four target words using a corpus-based and computational approach.

(2) Second contribution: I used more than two corpora. In order to collect a large amount of data, I used the Chinese Gigaword Corpus. In order to assign all possible appropriate concepts of word senses of the four target words, I employed HowNet as my knowledge base. In order to evaluate their performances, I used both Chinese Wordnet (CWN) and Xiandai Hanyu Cidian (Xian Han). By utilizing these four corpora, I was able to explore and deal with different conditions and problems and integrate them for the four target words in this sense prediction study.

(3) Third contribution: I ran off-line multiple-choice tasks of the experimental evaluations. The main goal of this study was to predict all possible sense for the four target words using a corpus-driven linguistic approach. In order to demonstrate the accuracy rates of these performances by automatic programming in the corpus-based and computational approach, I used off-line multiple-choice tasks; in doing so, I was able to examine words that were the same in concept and regard them as having the same sense via native speakers' intuitions.