Automatic Discovery of Lexical Semantic Differences between Hong Kong Written Chinese and Standard Chinese with Computational Models

Yuanbing Zhao, Hongzhi Xu

Institute of Corpus Studies and Applications, Shanghai International Studies University

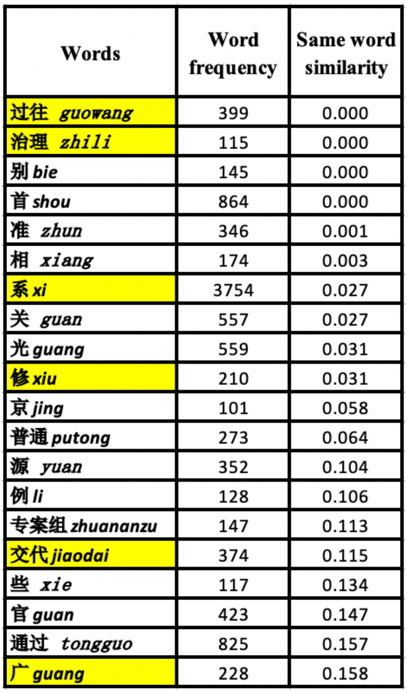
joyvinab@shisu.edu.cn, hxu@shisu.edu.cn

This study explores the possibility of utilizing computational models to automatically identify lexical semantic differences between Hong Kong Written Chinese (HKWC) and Standard Chinese (SC). In particular, we adopt the word embeddings technique to train word vectors and to examine a kind of lexical variation: homophone (same orthographic forms represent different meanings), which can be reflected by two words of the same form from two different origins showing a big distance in the space.

In order to train word embeddings for the two origins, we create a comparative corpus by collecting news articles from China News Network (SC) and Oriental Daily News (HKWC)'s society sections ranging from 2015 to 2019. Each corpus contains over 25 million characters. Then the two vector spaces are created and aligned.

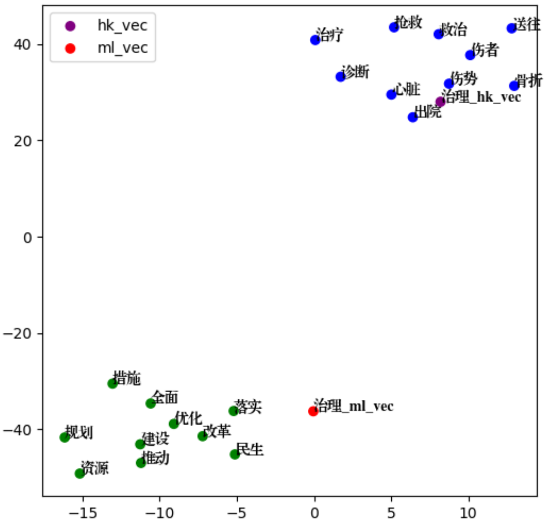
Since the data collection is still in progress, an initial experiment is conducted on a smaller corpus containing news articles of 2015. Firstly, we measure the similarities of words with the same forms. We find that the average similarities of the same word forms of approximately 0.68 between corresponding terms in both corpora, with a maximum of 0.93 and a minimum of 0.0 (we cast all negative similarities to zero).

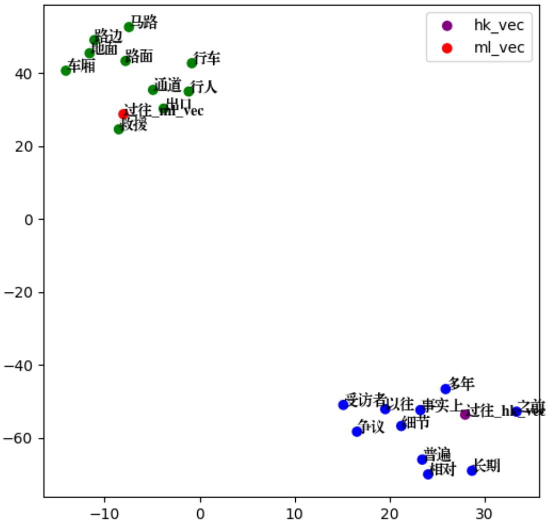
The table below lists the 20 least similar words. Among them, the six words that are emphasized might be homonyms with actual lexical distinctions, while the remaining low similarities might be the result of corpus dispersion, insufficient corpus volume, etc.



Words	Word frequency	Same word similarity
过往 *guowang*	399	0.000
治理 *zhili*	115	0.000
别 *bie*	145	0.000
首 *shou*	864	0.000
准 *zhun*	346	0.001
相 *xiang*	174	0.003
系 *xi*	3754	0.027
关 *guan*	557	0.027
光 *guang*	559	0.031
修 *xiu*	210	0.031
京 *jing*	101	0.058
普通 *putong*	273	0.064
源 *yuan*	352	0.104
例 *li*	128	0.106
专案组 *zhuananzu*	147	0.113
交代 *jiaodai*	374	0.115
些 *xie*	117	0.134
官 *guan*	423	0.147
通过 *tongguo*	825	0.157
广 *guang*	228	0.158

By comparing words with high similarity of HK vectors and the Mainland vectors respectively, we can get scatter plots similar to the following for each word. Words with semantic distinctions mentioned in previous research, like "单位" *danwei* and "牌照" *paizhao*, also show large semantic difference in our findings. Furthermore, there are interesting findings that are not discussed before. For example, the word "治理" in Hong Kong refers to "relief," while in Mainland it denotes "rectification"; the word "过往" in HK refers to time, whereas in Mainland, it frequently refers to the act of "passing through."





These are the preliminary results of our experiment. Next, we will incorporate collocation-based cluster analysis to conduct more thorough research and look for both homophones and allomorphs (different orthographic forms representing the same meanings) between two areas.

References

[1] 石定栩, 邵敬敏, and 朱志瑜. "港式中文与标准中文的比较 (第二版)." 香港: 香港教育图书公司 (2014).

[2] 张明辉, and 张静. "香港与内地书面汉语对比研究综述." 辽东学院学报 (社会科学版) 18.6 (2016): 91-97.