A New and Improved Method to Text-Mining in Chinese: Closer Language Segmentation in Detecting the Shifting Meaning of Patriotism

Annie S. Chao
mrsannechao@gmail.com
Rice University, United States of America

Qiwei Li
liqiwei2000@gmail.com
Rice University, United States of America

We aim to demonstrate a new methodology for detecting shifting nuances of ideas in Chinese intellectual history. Our subject is Chen Duxiu (1879-1942), founder of the Chinese Communist Party (CCP) and one of the most important historical figures of twentieth century China. We combine the latest text-mining tools with statistical analysis and natural language law, based on word frequency calculations. Because the Chinese language does not have spaces between words, and because each word is comprised of either a single or multiple characters or morphemes, segmentation of Chinese text poses a set of different challenges than for English corpus. By using a Chinese tokenization plug-in for R called JiebaR, we have developed a more precise method to create a Chinese natural language curve that conforms closely to Zipf's law, a commonly used model for distribution of words in a corpus. Zipf's law states that given a body of natural language text, the frequency of any word is inversely proportional to its rank. (Ha et al., 2003) For Chinese language, Zipf's law applies for words made up of multiple morphemes. (Xiao, 2008)

Our assumption is that a word is significant when its position on our curve deviates from the fitted curve based on Zipf's law. Departing from an existing method developed by Prof. Jin's team, we created two groups of keywords, and called them "anchor" and "companion" words. (Jin et al., 2014) "Anchor" words have large residuals (high deviation from the standard curve), and "companion" words have a high correlation with "anchor" words. We used the formula for Pearson's correlation coefficient to find the companion words. The coefficient has a value of between +1 and -1. The companion words provide the context with which to interpret the anchor words. Unlike Prof. Jin's team, we included keywords made up of more than two characters, such as the three character word for Nationalist party, Guomindang (國民黨), the four character word for citizen's assembly, guomin huiyi(國民會議), and the five character word for Marxism, makesizhuyi (馬克思主義).

The goal of our research is to analyze changing meaning of the concept of patriotism in Chen's writing from the beginning of his publishing career, ca. 1897, to the end of his life, 1942. Why is Chen such a fascinating figure to study? In addition to creating a political party that changed the course of Chinese history, Chen was equally, if not more, influential in bringing about the first cultural revolution of twentieth century China. Living at a time of great political unrest, Chen wrote passionately on the need to reform the people by revolutionizing Chinese culture, thoughts, and politics. After founding the CCP in 1920-21, Chen was expelled from the party in 1929 due to ideological differences. He was subsequently jailed by Chiang Kai-shek's Nationalist Party, and died a political pariah a few years later. As his political and personal fortune waxed and waned, Chen's conception of patriotism shifted. For this paper, we chose six important anchor words: citizen (國民), youth (青年), democracy (民主), revolution (革命), people (民族) and being patriotic (愛國). Our method is replicable for other corpus, with the understanding that the conclusion is derived with the additional layer of human deliberation.

Bibliography

Ha, L.Q., Sicilia-Garcia, E.I., Ming, J., and Smith, F.J. (2003). Extension of Zipf's Law to Word and Character N-grams for English and Chinese, *Computational Linguistics and Chinese Language Processing,* 8:1, 77-102.

Jin, G., Leong, Y., Yu, Y., and Liu, C., (2014). Application of Statistical Residual Analysis to Humanities Studies: Using Xin Qing Nian as Example, *Journal of the History of Ideas in East Asia*, 6: 327-366.

Xiao, H., (2008). On the Applicability of Zipf's Law in Chinese Word Frequency Distribution, *Journal of Chinese Language and Computing*, 18:1, 33-46.