Developing Simplified Chinese Psychological Linguistic Analysis Dictionary for Microblog

Rui Gao[1], Bibo Hao[1], He Li[2], Yusong Gao[1], and Tingshao Zhu[1]

[1] Institute of Psychology, University of Chinese Academy of Sciences
Chinese Academy of Sciences, Beijing 100190, P.R. China
[2] National Computer System Engineering Research Institute of China
Beijing, 100083, P.R. China
tszhu@psych.ac.cn,
{gaorui11,haobibo12}@mails.ucas.ac.cn

Abstract. The words that people use could reveal their emotional states, intentions, thinking styles, individual differences, etc. LIWC (Linguistic Inquiry and Word Count) has been widely used for psychological text analysis, and its dictionary is the core. The Traditional Chinese version of LIWC dictionary has been released, which is a translation of LIWC English dictionary. However, Simplified Chinese which is the world's most widely used language has subtle differences with Traditional Chinese. Furthermore, both English LIWC dictionary and Traditional Chinese version dictionary were both developed for relatively formal text. Microblog has become more and more popular in China nowadays. Original LIWC dictionaries take less consideration on microblog popular words, which makes it less applicable for text analysis on microblog. In this study, a Simplified Chinese LIWC dictionary is established according to LIWC categories. After translating Traditional Chinese dictionary into Simplified Chinese, five thousand words most frequently used in microblog are added into the dictionary. Four graduate students of psychology rated whether each word belonged in a category. The reliability and validity of Simplified Chinese LIWC dictionary were tested by these four judges. This new dictionary could contribute to all the text analysis on microblog in future.

Keywords: LIWC, Traditional Chinese, Simplified Chinese, microblog, text analysis.

1 Introduction

The rapid developing social media--microblog has had a significant impact on society, politics, economy, culture and people's daily life [1, 2]. Researchers have carried out a number of studies on microblog [3-7]. Computerized text analysis methods like LIWC (Linguistic Inquiry and Word Count) [8, 9] have been widely used for social media researches [2, 10-12]. LIWC dictionary is the core of LIWC text analysis method [8, 9, 13].

K. Imamura et al. (Eds.): BHI 2013, LNAI 8211, pp. 359–368, 2013.

Simplified Chinese now is the world's most widely used language, but it cannot be analyzed with LIWC because of the vacancy of Simplified Chinese version of dictionary. The Traditional Chinese version of LIWC dictionary — CLIWC(Chinese Linguistic Inquiry and Word Count) [14] dictionary has been released, which makes it possible to analyze Traditional Chinese text with LIWC software. But, Simplified Chinese has subtle differences with Traditional Chinese. Furthermore, both English LIWC dictionary and CLIWC dictionary were both developed for relatively formal text.

In this study, specific exclusive Simplified Chinese LIWC dictionary (SCLIWC) was established according to LIWC dictionary and CLIWC dictionary, and then microblog high frequency words were added into SCLIWC. This dictionary, SCMBWC (Simplified Chinese Microblog Word Count) is a promising approach for both psychological and other kinds of researches based on Microblog.

The rest of this paper is organized as follows. In Section 2, we overview some related work. Section 3 describes how to build the dictionary. The experimental results and discussion are presented in Section 4, followed by the conclusion and future work in Section 5.

2 Related Work

LIWC with its English dictionary is one of the most prestigious tools of content analysis [15]. First significant version of LIWC was released in 1997, after continuing optimizing for decade the latest version of LIWC software and English dictionary is LIWC2007 [9]. LIWC is a milestone in the history of computerized text analysis, and plenty of researches are based on LIWC [16-20].

Establishment of CLIWC made it possible to use computerized text analysis methods in Traditional Chinese text analysis related researches. CLIWC has made an outstanding contribution to Traditional Chinese content analysis area [14].

Traditional Chinese and Chinese Simplified share the same origin; however, along with the development of the times, diversity has been evolved between them [21]. Many Traditional Chinese words, cannot find a unique identifying Chinese Simplified word correspond with it. Figure 1 shows some examples of this kind of words. Furthermore, words spelled the same in these two languages might express dissimilar meanings [22, 23]. More crucial is, compared to differences of the two languages itself, linguistic using differences in their populations merited to be taken into serious consideration [13, 21, 24].

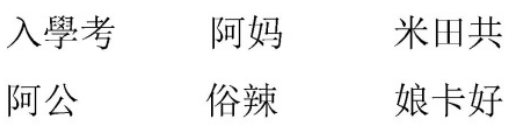

Fig. 1. Examples of Word could not find unique corresponding Chinese Simplified word