Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 47 (2018) 194–213

www.elsevier.com/locate/csl

Using Chinese radical parts for sentiment analysis and domain-dependent seed set extraction☆

August F.Y. Chao, Heng-Li Yang*

Department of Management Information Systems, National Cheng-Chi University, 64, Sec. 2, Chihnan Road, Wenshan District, Taipei, Taiwan

Received 3 November 2015; received in revised form 19 June 2017; accepted 24 July 2017
Available online 7 August 2017

Abstract

Although there has been good progress in English sentiment analysis and resources, studies in English cannot be directly used in Chinese owing to the nature of Chinese language. Previous studies suggested adopting linguistic information, such as grammar and morpheme information, to assist in sentiment analysis for Chinese text. However, morpheme-based approaches have a problem in identifying seeds. In addition, these methods do not take advantage of radicals in the characters, which contain a great deal of semantic information. A Chinese word is composed of one or more characters, each of which has its radical part. We can interpret the partial meaning of a character by analyzing that of the radical in the character. Therefore, we not only consider the radical information as the semantic root of a character, but also consider the radical parts between characters in a word as an appropriate linguistic unit for conducting sentiment analysis.

In this study, we conducted a series of experiments using radicals as the feature unit in sentiment analysis. Using segmented results from part-of-speech tools as a meaningful linguistic unit (word) in Chinese, we conducted analyses of single-feature word (unigram) and frequently seen two words (pointwise mutual information collocated bigrams) through various sentiment analysis measures. It is concluded that radical features could work better than word features and would consume less computing memory and time. An extended study of the extraction of seeds was also conducted, and the results indicated that 50 seed radical features performed well. A cross-corpus comparison was also conducted; the results demonstrated that the use of 50 extracted radical features as domain-dependent keywords worked better than other sentiment analysis strategies. This study confirmed that radical information could be adopted as a feature unit in sentiment analysis and that domain-dependent radicals could be reused in different corpora.

Keywords: Sentiment analysis; Chinese radical; Restaurant review analysis; Domain-dependent seed

1. Introduction

Reviews are central to almost all human activities and are key influencers of our behavior (Liu, 2012). From the perspective of business owners, monitoring online comments has become an important marketing strategy for

☆ This paper has been recommended for acceptance by S. Narayanan.

* Corresponding author.
E-mail address: aug.chao@gmail.com (A.F.Y. Chao), yanh@nccu.edu.tw (H.-L. Yang).

http://dx.doi.org/10.1016/j.csl.2017.07.007
0885-2308/

understanding customers. From the perspective of consumers, it is crucial to be able to gain an integrated and comprehensive understanding of available services and products. Thus, we need sentiment analysis of reviews when inundated with a huge array of comments on the Internet. However, sentiment analysis faces many challenges. First, user-generated content is presented in an unstructured format and contains more details than Likert-style survey responses (Pan et al., 2007); therefore, it is difficult to establish a fixed model for these different types of contents. Second, before formulating review analysis patterns, we require knowledge of the different aspects of a product that might be of concern to the customers. For example, in restaurant reviews, the amount of time spent waiting and queuing is relevant to consumers. Language resources used for sentiment analysis are domain dependent (Pang and Lee, 2008); therefore, it is difficult to create a universal sentiment lexicon for general purposes. People use their native language to describe their experiences and express their sentiments; thus, comprehending the semantic meaning of written reviews requires massive natural language processing (NLP) and additional dependent language resources. These challenges have driven the rapid development in sentiment analysis and NLP studies in recent years, and many language resources have been introduced to facilitate comprehension of sentiments in written texts.

Unfortunately, not all language resources and techniques can be directly adapted to the Chinese language environment for text mining (Yang and Chao, 2015; Zheng et al., 2015). Chinese is a non-space-delimited, polysyllabic, and sequentially interpreted language (Chao, 1965), which has its own characters, lexicons, and unique sentence grammar, and these characteristics present language-dependent problems in the development of sentiment analysis. To seek a native approach for analyzing Chinese reviews, researchers have used linguistic information, such as morphological words (Ku et al., 2009; Liu, 2010; Lu et al., 2010; Fu and Wang, 2010; Zhang et al., 2012; Yang and Chao, 2015) as a leverage to uncover the semantic meaning of reviews.

Chinese morphological words, which can consist of more than one character, have easily observable semantic meanings and can be used as a guideline for classifying words or doing other in-depth analysis (Ku et al., 2009). Several studies have shown that morphological information can assist in the identification of sentiments at lexicon level (Ku et al., 2006; Liu et al., 2010), aspect-sentiment level (Zhang et al., 2012), sentence level (Ku et al., 2009; Fu and Wang, 2010), and document level (Yang and Chao, 2015). However, in morpheme-based approaches, providing the initial morpheme words (i. e., seeds) is a challenging task and needs domain experts.

In addition to morphological information, it can be observed that Chinese morpheme characters are logograms and that almost 90% of Chinese characters can be disassembled into semantic radicals (“義符”) and phonetic radicals (“音符”) (Li and Kang, 1993). These semantic symbols can generally be used as a radical index for organizing characters, and they can reveal the basic concepts mixed within a single character (Huang et al., 2008). Such an observation motivates us in this study to try to use radical information to relieve the burden of locating morpheme words in the morpheme-based approach. We consider the radical parts (i.e., a radical form of two or more feature words) to be the basic concept compositions, which can be used to identify sentiment features at the linguistic conceptual level. For example, the words “早餐” (breakfast) and “晚餐” (dinner) can be considered as a combination of the radical “日” (day, time) and the radical “食” (food), as well as be comprehended as “food for a specific time.” Another example, words with a negative meaning, e.g., “憤怒” (angry) and “懶惰” (lazy), can be considered as a combination of the radical “忄” (heart) and the radical “忄” (heart), as well as be comprehended as “feeling from the heart.”

To our knowledge, prior research has not formally investigated the use of radical information to facilitate sentiment analysis. The purpose of this study was to demonstrate the superiority of applying radical information. Since previous studies have shown that morpheme-based sentiment analysis approaches can be more effective than text-retrieval and keyword-based approaches (Jang and Shin, 2010; Yang and Chao, 2015), specifically, this study tries to demonstrate the advantages of the radical-based approach over morpheme word-based approaches. We tried two cases of comparison: single-feature word (unigram) and two frequently seen words (i.e., pointwise mutual information collocated bigrams). In the case of bigrams, the morpheme roots were applied. In both cases, we compared a word-based approach with its corresponding radical-based approach. Finally, we propose that the extracted radical features be used as domain-dependent keywords to analyze similar domain reviews from different sources. In all of the above cases, we also made a subsidiary comparison—comparing with a keyword-based approach. Since our collected review corpora are in traditional Chinese, those language resources (e.g., HowNet) in simplified Chinese are not suitable for this study because there are different cultural problems and different radicals for the same character. For example, the word 鳳 (phoenix) with the radical 鳥 (bird) in traditional Chinese becomes the word 凤 with the radical 几 (small table) in simplified Chinese. There are some resources for traditional Chinese, e.g., Chinese WordNet,