# A Survey of Automatic Acquisition of Sentiment Lexicon Focusing on Relationship between Approach and Domain

2130409 Terumasa Taka

Sentiment analysis is a task to classify a writer's opinion expressed in a text into positive, negative, or neutral. It is utilized for various purposes including marketing and investment. A common resource for sentiment analysis is a sentiment lexicon, a database that collects sentiment words/phrases and defines their scores of polarity (positive, negative, or neutral). The polarity of a text can be determined as positive or negative when a polarity score of the entire text is greater or less than zero, where the score of the text is calculated based on the polarity scores of the individual words in the text. However, since the polarity of a word can be changed for different domains of texts, the sentiment analysis using a general (domain-independent) sentiment lexicon may fail. Although it is preferrable to use a domain-specific sentiment lexicon, it takes a lot of time and efforts to construct it manually. Therefore, automatic acquisition of high-quality domain-specific sentiment lexicons from domain-specific corpora has attracted much attention.

There are two approaches to the automatic acquisition of sentiment lexicons: lexicon construction, in which a new domain-specific sentiment lexicon is constructed from scratch, and lexicon adaptation, in which an existing sentiment lexicon is adapted to a specific domain. On the one hand, lexicon construction starts with acquiring sentiment words, including words used only in a specific domain. On the other hand, no new sentiment word is acquired in lexicon adaptation, but the polarity of sentiment words and their scores are modified for a specific domain.

In previous surveys on sentiment analysis and automatic acquisition of sentiment lexicons, existing studies have been summarized in terms of technologies, data sources, and target languages. However, to the best our knowledge, no survey has focused on the relationship between the approach to the automatic acquisition of sentiment lexicons (i.e., lexicon construction or lexicon adaptation) and the domain of a target text.

This research investigates a trend of research on automatic acquisition of sentiment lexicons, focusing on the approach of lexicon acquisition and the domain of a text. Two approaches are considered, i.e., lexicon construction and lexicon adaptation, while six domains are considered, i.e., political speech, news, movie reviews, product reviews, social media, and others. Technical papers about automatic acquisition of sentiment lexicons are collected by searching Web with several keywords by Google Scholar and then manually selecting the genuine related papers from the top ranked ones. Next, a

matrix of the approaches by the domains is prepared. The approach and the target domain of a method proposed in each paper is manually identified, then the paper is fit into one of the cells in the matrix. We investigate what pairs of the approach and the domain for which many papers are fit in order to reveal an overall trend of the research. In addition, we discuss how the target domain influences researchers' choice of the approach of lexicon acquisition from individual case studies.

After searching and manually selecting papers, 417 papers were investigated in this survey. The number of papers for each pair of (approach, domain) is as follows. As for lexicon construction, 3 papers were found for (lexicon construction, political speech), 26 for (lexicon construction, news), 33 for (lexicon construction, movie reviews), 112 for (lexicon construction, product reviews), 133 for (lexicon construction, social media), and 55 for (lexicon construction, others). As for lexicon adaptation, 2 papers were found for (lexicon adaptation, political speech), 2 for (lexicon adaptation, news), 11 for (lexicon adaptation, movie review), 25 for (lexicon adaptation, product review), 4 for (lexicon adaptation, social media), and 11 for (lexicon adaptation, others).

Comparing the number of papers of lexicon construction and lexicon adaptation, it was found that lexicon construction accounted for 86.8% of the total number of papers and lexicon adaptation did for 13.2%, indicating that lexicon construction is the major approach. Seeing the proportion of two approaches in individual domains, 97.1% of the papers in the social media domain and 92.9% in the news domain took the lexicon construction approach. This may be because new words are often used in social media and news, and lexicon construction, which starts with collecting sentiment words from texts, is more appropriate. Besides, political speech (40%), movie reviews (25.0%), and product reviews (18.2%) were the domains where the lexicon adaptation approach was relatively more frequently employed. It is supposed that lexicon adaptation is appropriate for the domain of political speech, where new words are less likely to appear. Before the survey, we expected that lexicon construction would be suitable for movie and product reviews because domain-specific words were often used in these reviews. However, lexicon adaptation was also used to some extent in the movie and product review domains. In the studies for the “other” domain, comments on live videos, texts in video games or Chinese poetry was analyzed. Sentiment lexicons for these target texts were often acquired by lexicon construction method to analyze contexts specific to the domain and words/expressions used in a specific community. This may be because it is necessary to automatically collect domain-specific sentiment words to precisely capture the unique language, expression, and cultural context of the domain.

Comparing the number of papers in the individual domains, studies in the product

review domain and the social media domain accounted for the most significant number of papers. The proportion of each domain was 32.9%. When limited to papers of lexicon construction, the social media domain had the most significant number of papers (36.7%), supporting the aforementioned discussion that lexicon construction methods are more likely to be used in social media where there are many new words. However, even for the social domain, the approach of lexicon adaptation is also promising when a general sentiment lexicon that compiles many new sentiment words is available. Besides, when limited to papers of lexicon adaptation, it is frequently applied for the product review domain (45.5%) and the movie review domain (20.0%). It means that there are relatively more studies that utilize a sentiment lexicon adopted to the target domain for the sentiment analysis of review texts.

From the above findings, it was confirmed that the approach to automatic acquisition of sentiment lexicons is closely related to the domain for which they are applied. Selecting and applying the most appropriate approach according to the characteristics and needs of each domain is the key to improving the accuracy and effectiveness of sentiment analysis.