Expert Systems With Applications 89 (2017) 273–285

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A two-fold rule-based model for aspect extraction

Toqir A. Rana*, Yu-N Cheah

School of Computer Sciences, Universiti Sains Malaysia, Malaysia

ARTICLE INFO

Article history:
Received 17 May 2017
Revised 26 July 2017
Accepted 27 July 2017
Available online 28 July 2017

Keywords:
Aspect-based sentiment analysis
Opinion mining
Aspect extraction
Explicit aspects
Sequential pattern-based rules
Aspect pruning

ABSTRACT

Opinion target extraction or aspect extraction is the most important subtask of the aspect-based sentiment analysis. This task focuses on the identification of the targets of user's opinions or sentiments from online reviews. In the recent years, syntactic patterns-based approaches have performed quite well and produced significant improvement in the aspect extraction task. However, these approaches are heavily dependent on the dependency parsers which produced syntactic relations following the grammatical rules and language constraints. In contemporary, users do not give much importance to these rules and constraints while expressing their opinions about particular product and neither reviewer websites restrict users to do so. This makes syntactic patterns-based approaches vulnerable. Therefore, in this paper, we are proposing a two-fold rules-based model (TF-RBM) which uses rules defined on the basis of sequential patterns mined from customer reviews. The first fold extracts aspects associated with domain independent opinions and the second fold extracts aspects associated with domain dependent opinions. We have also applied frequency- and similarity-based approaches to improve the aspect extraction accuracy of the proposed model. Our experimental evaluation has shown better results as compared with the state-of-the-art and most recent approaches.

1. Introduction

Continuous expansion of World Wide Web (WWW) and tremendous growth of social media networks have rehabilitated the life style of a common person. Users of the Internet in general, and social media networks in particular, are turning to these platforms to share their experiences, emotions and feelings about different events, places or products in the form of reviews. In other words, the Internet and social media have become the prime source for decision making and information gathering rather than conducting traditional surveys. People are increasingly relying on the experiences of other users to buy new product, to travel, or to choose an appropriate hotel. However, it is almost impossible for any user to read all such reviews and make a right decision. Therefore, sentiment analysis plays a vital role in analyzing these reviews and producing an overall summary of these reviews (Cambria, 2016).

Among the different granularity levels of sentiment analysis, aspect-based sentiment analysis has attracted a large number of researchers (Liu, 2012; Pang & Lee, 2008). Aspect-based sentiment analysis deals with the users' sentiments or opinions and the targets of these opinions which are often referred to as aspects, and generates an overall summary of these aspects along with the positive or negative polarity of users' opinions towards that specific aspect. Among different tasks of aspect-based sentiment analysis, aspect extraction is the most important task and studied by a large number of researchers (Rana & Cheah, 2016a). This task involves how to identify aspects which are related to the specific entity and what are the users' opinions which are related to those aspects.

Hu and Liu (2004) have identified two types of aspects: explicit and implicit. Explicit aspects are those aspects which are expressed explicitly. For example, in the sentence: "The phone is great", "phone" is the aspect. While in this sentence: "Phone is small", the user is talking about the "size" of the phone but did not mentioned it explicitly in this sentence, and such aspects are called implicit aspects. Most of the researchers have focused on explicit aspects and only a very small number of studies have focused on the implicit aspects (Rana & Cheah, 2016a).

In recent years, linguistic patterns-based approaches have been widely studied for the extraction of explicit aspects (Bancken, Alfarone, & Davis, 2014; Du, Chan, & Zhou, 2014; Kang & Zhou, 2016; Liu, Gao, Liu, & Zhang, 2015, Liu, Gao, Liu, & Zhang, 2016; Liu, Liu, Zhang, Kim, & Gao, 2016; Poria, Cambria, Ku, Gui, & Gelbukh, 2014; Qiu, Liu, Bu, & Chen, 2011). Most of these studies used dependency parser-based approaches to explore relations among aspects and opinions. These dependency parsers heavily depend on the grammatical rules and language constraints. On the other hand, reviewer websites do not restrict users to follow these rules and

* Corresponding author.
E-mail addresses: toqirr@gmail.com (T.A. Rana), yncheah@usm.my (Y.-N. Cheah).

http://dx.doi.org/10.1016/j.eswa.2017.07.047
0957-4174/

therefore, users tend to write in casual manner and sometimes violate such rules. For example: “same story as everybody else when trying to get service from apex - nothing”. In this sentence, “service” is the aspect which is associated with the opinion word “nothing”, but dependency parsers are unable to relate them. Machine learning approaches, like conditional random fields (CRF), have also been applied to solve the problem (Chen, Qi, & Wang, 2012; Choi & Cardie, 2010; Huang, Liu, Peng, & Niu, 2012; Jakob & Gurevych, 2010; Yang & Cardie, 2013). However, CRF-based approaches are supervised and required a large number of aspects to perform well.

In this paper, we have proposed a two-fold rule-based model (TF-RBM) which performs the task in three steps; (1) using sequential patterns-based rules (SPR) (Rana & Cheah, 2015, 2016c) for the extraction of explicit aspects which are associated with regular opinions; (2) improving aspect extraction accuracy with a frequency-based approach along with normalized Google distance (NGD) (Cilibrasi & Vitanyi, 2007); and (3) in the second fold, extracting aspects which are associated with domain dependent opinions, we called this phase as concept extraction. Although, WordNet (Fellbaum, 1998) and Word2Vec[1] similarity measures have been used by Kang and Zhou (2016) and Liu et al. (2016) respectively, but both of these tools require a huge corpus and trained model. On the other hand, NGD does not require any corpus or trained model. NGD uses Google as the search engine and calculates the similarity between two terms on the basis of the hits returned by Google.

The proposed approach first uses SPRs to extract all aspects and opinions. Although, a large number of aspects can be extracted, many are not related to the product. To overcome this issue, pruning is carried out by calculating the frequency of each word and eliminating all aspects which do not fulfill a minimum threshold. However, not all the non-frequent aspects are irrelevant. To find the non-frequent, but relevant, aspects, we calculate the NGD score for only those aspects which are nouns and select those which have NGDs lower than the minimum score, even though they are not frequent as proposed in Rana and Cheah (2018). In the second fold, we search for aspects which are not associated with any regular opinion, but are associated with some domain dependent opinions e.g., small, large, tiny. We have compared our approach with the state-of-the-art and most recent approaches. Experimental results of the proposed model have shown the improvement in the aspect extraction phase.

The remaining of the paper is organized as follows: Section 2 outlines the related work. In Section 3, the proposed methodology for aspect extraction is discussed. In Section 4, the experimental results are presented while Section 5 concludes the paper.

2. Related work

2.1. Unsupervised approaches

Aspect extraction in aspect-based sentiment analysis was first studied by Hu and Liu (2004). They considered nouns as potential aspects and calculated the frequency of each aspect. They used association rule miner (CBA) (Liu, Hsu, & Ma, 1998) for the generation of candidate aspects. CBA generates frequent noun phrases without considering the position of each noun in the sentence. Therefore, they used aspect pruning to eliminate aspect phrases where the noun words were not closely related. Furthermore, the frequency of each aspect was calculated and only those aspects were selected which had the support greater than minimum given threshold. For each selected aspect, the nearest adjective was considered as the opinion and these opinions were further used to identify the non-frequent aspects. Popescu and Etzioni (2007) later improved the accuracy by calculating Pointwise Mutual Information (PMI) for each aspect. If the respective calculated PMIs were too low, those aspects were eliminated. Li, Zhou, and Li (2015) proposed a web-based similarity method along with frequency pruning method to improve the aspect extraction accuracy. They adopted the same approach for frequency pruning as proposed by Hu and Liu. For web-based similarity, number of hits returned by a search engine, for entity term and aspect terms, were used to evaluate similarity among two terms. They calculated the PMI-IR score and those aspects were eliminated which have a score less than the given threshold.

Many researchers followed Hu and Liu’s work. Li, Zhang, Ma, Zhou, and Sun (2009) combined NLP with statistical methods to identify aspects from Chinese online reviews. Raju, Pingali, and Varma (2009) proposed a three step clustering-based approach to identify product aspects. Their approach grouped similar nouns/noun phrases into clusters and these clusters were used to identify aspects.

Reviewing websites provide additional information related to the customer reviews e.g., rating. Such additional information was collected from the review sites to identify product aspects (Meng & Wang, 2009; Moghaddam & Ester, 2010). Eirinaki, Pisal, and Singh (2012) also extracted aspects considering the most frequent aspects and Marrese-Taylor, Velásquez, and Bravo-Marquez (2014) also adopt a similar approach for the tourism domain. Other than these approaches, bootstrapping techniques were explored by Bagheri, Saraee, and De Jong (2013), Li, Qin, et al. (2015) and Zhu, Wang, Zhu, Tsou, and Ma (2011). Gunes, Furche, Shoreditch, and Orsi (2016) proposed a frequency-based noun-clustering method to detect structured aspect terms. Quan and Ren (2014) used PMI along with document frequency to explore the association between aspects and opinions. Yang, Liu, Lin, and Lin (2016) combined local context information, i.e., within a sentence, and global context information, i.e., within multiple sentences in a document, for aspect extraction and ranked them on the basis of their score and frequency.

2.2. Semi-supervised approaches

Wang and Wang (2008) and Hai, Chang, and Cong (2012) used a bootstrapping method to learn both product aspects and opinions from Chinese customer reviews. Wei, Chen, Yang, and Yang (2010) followed a semantic-based approach which identified the most frequent aspects and further eliminated all aspects which were irrelevant to the seed opinion words. Ma, Zhang, Yan, and Kim (2013) combined Latent Dirichlet Allocation (LDA) with a lexicon of synonyms to extract aspects from Chinese reviews. Liu, Xu, et al. (2015) proposed a word alignment-based model which identified aspects from the position of the words in the sentence. Yan, Xing, Zhang, and Ma (2015) used association among aspects and opinion words and used a synonym lexicon to expand the aspect list. Samha, Li, and Zhang (2014) built a list of aspects, from the information provided by the manufactures, to identify similar aspects from customer reviews.

2.3. Supervised approaches

A dictionary-based supervised approach was proposed by Kobayashi, Inui, and Matsumoto (2007) to identify aspects and syntactic patterns were used to find associations among aspects, opinions and products. Cruz, Troyano, Enríquez, Ortega, and Vallejo (2013) generated a taxonomy for aspects, specific for every domain, and validated the opinions by the assumption that opinions appeared somewhere near to aspects in the sentence.

[1] https://code.google.com/p/word2vec/