Information Sciences 272 (2014) 16–28

Contents lists available at ScienceDirect

Information Sciences

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

Unsupervised product feature extraction for feature-oriented opinion determination

Changqin Quan [a,*], Fuji Ren [a,b]

[a] AnHui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine, School of Computer and Information, HeFei University of Technology, Hefei 230009, China
[b] Faculty of Engineering, University of Tokushima, 2-1 Minami-Josanjima, Tokushima 770-8506, Japan

ARTICLE INFO

Article history:
Received 5 March 2013
Received in revised form 20 January 2014
Accepted 9 February 2014
Available online 19 February 2014

Keywords:
Product feature extraction
Sentiment analysis
Domain corpora
Term similarity
Opinion lexicon

ABSTRACT

Identifying product features from reviews is the fundamental step as well as a bottleneck in feature-level sentiment analysis. This study proposes a method of unsupervised product feature extraction for feature-oriented opinion determination. The domain-specific features are extracted by measuring the similarity distance of domain vectors. A domain vector is derived based on the association values between a feature and comparative domain corpora. A novel term similarity measure (PMI–TFIDF) is introduced to evaluate the association of candidate features and domain entities. The results show that our approach of feature extraction outperforms other state-of-the-art methods, and the only external resources used are comparative domain corpora. Therefore, it is generic and unsupervised. Compared with traditional pointwise mutual information (PMI), PMI–TFIDF showed better distinction ability. We also propose feature-oriented opinion determination based on feature-opinion pair extraction and feature-oriented opinion lexicon generation. The results demonstrate the effectiveness of our proposed method and indicate that feature-oriented opinion lexicons are superior to general opinion lexicons for feature-oriented opinion determination.

1. Introduction

With the growing popularity and availability of user-generated reviews, recently increasing more attention is being paid to the research of sentiment analysis (SA). SA is also called opinion mining, and aims at determining the attitude polarity (positive, negative or neutral) of a text. The potential benefits of SA on education, e-commerce, and opinion polls are significant. Consequently, a variety of techniques have performed to probe SA. SA techniques basically include two branches: document-level SA and feature-level SA. Document-level SA focuses on determining an overall opinion for a document. Representative work includes Turney applied an unsupervised method for detecting the polarity of product reviews [46]; Pang applied machine learning techniques for sentiment classification of movie reviews [29]. The following researches on predicting star ratings on different scales extended the basic task of classifying a review as either positive or negative [30,35,39,53].

Feature-level SA aims to extract features (e.g., camera's image quality, size) from reviews, and then determine opinions that are linked with each feature. Compared with document-level SA, feature-level SA is able to provide more fine grained

* Corresponding author. Tel.: +86 13003006323.
E-mail addresses: quanchqin@hfut.edu.cn (C. Quan), ren@is.tokushima-u.ac.jp (F. Ren).

http://dx.doi.org/10.1016/j.ins.2014.02.063
0020-0255/

sentiment analysis on certain opinion targets, and has a wider range of applications [24,49]. Representative work in feature-level SA includes Hu and Liu proposed a method for mining and summarizing product reviews [17]; Popescu and Etzioni extended Hu and Liu's method through web search [31]; Ding et al. proposed a holistic lexicon-based approach to improve Hu and Liu's method [10].

Identifying product features from reviews is the fundamental step as well as a bottleneck in the problem of feature-level SA. Although several researchers have studied opinion lexicon expansion and opinion product feature extraction problems, their algorithms either need additional and external resources or impose strong constraints, and are of limited success [33].

As SA is sensitive to the domain, product feature extraction can be seen as a problem of domain-specific entity recognition. However, most of the existing methods toward domain-specific entity recognition often rely on domain-specific knowledge to improve system performance, but such knowledge is expensive to build and maintain for various domains, and a domain dependent method is hard to extend to other domains. Therefore, extracting domain-specific product feature in a generic and unsupervised manner would be desirable for feature-level SA.

In this paper, we propose an approach that exploits domain-specificity of words as a form of domain-knowledge for domain-specific feature extraction. The basic idea of our approach is to extract domain product features through the evaluation of their weights in different domains. Compared to prior work, our approach is generic and unsupervised. In evaluation, we compare the proposed method with several state-of-the-art methods (including unsupervised and semi-supervised) using a standard product review test collection. The experimental results demonstrate that our approach outperforms other unsupervised and semi-supervised methods.

Furthermore, we explore the problem of product feature opinion determination, which includes three main subtasks: (1) feature-opinion pair extraction; (2) feature-oriented opinion lexicon generation; and (3) feature-oriented opinion determination. We compare the results with different experimental setups, and the experiments confirm the effectiveness of our proposed feature-opinion pair extraction method, and indicate that feature-oriented opinion lexicons are superior to general opinion lexicons for feature-oriented opinion determination.

The remainder of this paper is organized as follows. In Section 2, a literature review is conducted on the studies of automatic product feature extraction and product feature opinion determination. Section 3 presents the proposed method for domain-specific product feature extraction using comparative domain corpora. Section 4 describes feature-oriented opinion determination. Section 5 presents the evaluation and discussion. Section 6 concludes this work with closing remarks and future directions.

2. Literature review

Feature-level SA is essentially a content based information extraction and classification problem involving the main steps of feature extraction, relation extraction, opinion classification, etc. This study focuses on two fundamental problems including product feature extraction and product feature opinion determination.

2.1. Product feature extraction

Previous studies have explored many different methods for this problem. In [17], a frequency-based approach was proposed to find frequent nouns and noun phrases, and then opinion words were used to extract infrequent aspects based on the idea of "opinions have targets". This idea was generalized to a double propagation method [33] and a dependency based method [34]. Popescu and Etzioni extended Hu and Liu's method by computing the PMI score between phrase and class-specified discriminators through a web search [31]. The necessary of web querying in this work has been pointed out a problem [5]. Ding and Liu further improved Hu and Liu's system by adding some rules to handle different kinds of sentence structures [9]. However, sentence structure rules are language dependent, so it is hard to be applied to other languages.

Jin et al. proposed a supervised learning method based on lexicalized Hidden Markov Models (L-HMMs) that integrates several linguistic features [19]. The authors demonstrated that their method is more effective and accurate than rule based method. The main problem is that SA is sensitive to the domain of training data. A classifier trained using reviews from one domain often performs poorly in another domain.

Scaffidi et al. applied a language model approach with the assumption that product features are mentioned more often in a product review than they are mentioned in generic English [38]. But the results may sensitive on corpus size.

Somasundaran and Wiebe used syntactic dependency for aspect and opinion extraction [40]. Ma and Wan exploited aspects extraction in the news title and contents [25]. Yu et al. explored important aspects extraction [51]. More survey can be found in [24].

As SA is sensitive to domains, product feature extraction can be seen as a problem of domain-specific entity recognition. Named entity recognition (NER) is a preprocessing tool to many natural language processing tasks, such as text summarization [2], machine translation, and document classification [1,36]. Traditional NER task has expanded beyond identifying people, location, and organization to book titles, email addresses, phone numbers, and protein names [28]. Recently there has been a surge of interest in extracting product attributes from online reviews. Recent work on product attribute extraction by [4] applied a Latent Dirichlet Allocation (LDA) model to identify different features of products from user reviews. Veselin and Cardie treated feature extraction as a topic coreference resolution problem, and they proposed to train a classifier to judge if