Information Processing Letters 116 (2016) 623–627

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

Combining local and global information for product feature extraction in opinion documents

Liang Yang [a,*], Bing Liu [b], Hongfei Lin [a], Yuan Lin [a]

[a] Department of Computer Science and Technology, Dalian University of Technology, Dalian, China
[b] Department of Computer Science, University of Illinois in Chicago, Chicago, USA

ARTICLE INFO

Article history:
Received 23 March 2015
Received in revised form 12 April 2016
Accepted 25 April 2016
Available online 26 April 2016
Communicated by W.-L. Hsu

Keywords:
Opinion mining
Feature extraction
Local context information
Global context information
Graph algorithms

ABSTRACT

Product feature (feature in brief) extraction is one of important tasks in opinion mining as it enables an opinion mining system to provide feature level opinions. Most existing feature extraction methods use only local context information (LCI) in a clause or a sentence (such as co-occurrence or dependency relation) for extraction. But global context information (GCI) is also helpful. In this paper, we propose a combined approach, which integrates LCI and GCI to extract and rank features based on feature score and frequency. Experimental evaluation shows that the combined approach does a good job. It outperforms the baseline extraction methods individually.

1. Introduction

With the rapid expansion of online commerce, millions of users contribute an enormous number of product reviews every day. In those reviews, people praise and criticize a variety of features of each product. Some popular products can get thousands of reviews. This makes it hard for a customer to read and to evaluate the features, and to make the decision whether to buy the product. Unfortunately, the problem of extracting features from a product review corpus and rank them is still far from being solved.

Feature based opinion mining or sentiment analysis [3,18,10] is an active research area. One task is to extract product features in online reviews. For example, in "*The picture of the camera is good*," "*picture*" is a feature of *the camera*. There have been many existing studies of the problem, e.g., Hu and Liu [3], Popescu and Etzioni [17], Scaffidi et al. [22] and so on.

Hu and Liu [3] proposed a frequency based method to extract product features. They also used opinion words to help further. Popescu and Etzioni [17] computed the pointwise mutual information (PMI) score between a phrase and some predefined class specific discriminators to address the same problem. Ku et al. [8] made use of the TF-IDF scheme considering terms at the document level and paragraph level to extract features. Along these lines, Zhuang et al. [27] proposed a dependency based technique for feature extraction. Kobayashi et al. [9] used a pattern based method to mine feature–evaluation pairs. Scaffidi et al. [22] compared the frequency of extracted frequent nouns and noun phrases in a review corpus with their occurrence rates in a generic English corpus to identify true features. In Wang and Wang [24], a bootstrapping method is proposed to extract features. Wu et al. [25] exploited shallow and deep parsing dependency. Qiu et al. [19] proposed a double propagation method, which utilized certain syntactic relations of opinion words and features, and propagated through both of them iteratively. Zhang and Liu [26] adopted the HITS algorithm [7] for feature extraction by constructing a bipartite graph using the relations between

* Corresponding author.
E-mail addresses: yangliang@mail.dlut.edu.cn (L. Yang), liub@cs.uic.edu (B. Liu), hflin@dlut.edu.cn (H. Lin), zhlin@dlut.edu.cn (Y. Lin).

http://dx.doi.org/10.1016/j.ipl.2016.04.009
0020-0190/

opinion words and possible features. Features act as authorities and opinion words as hubs in HITS. If a feature candidate has a high authority score, it should be a highly-relevant feature. Moghaddam and Ester [14] augmented the frequency-based approach with an additional pattern-based filter to remove some non-feature terms. Long et al. [11] extracted noun features based on frequency and information distance. Poria and Cambria [16] proposed a rule-based approach to aspect extraction from product reviews. Their method first finds the core feature words using the frequency method.

Rule-based and statistical methods are also widely used in general information extraction. Early extraction systems are mainly based on rules [20]. The most popular models in statistical methods are Hidden Markov Models (HMM) [21], Maximum Entropy Models (ME) [1] and Conditional Random Fields (CRF) [13]. CRF has been shown to be the most effective method. It has been used in [23,5,12,2,4]. However, CRF has a limitation that it only captures local patterns rather than long range patterns. Qiu et al. [19] showed that many features and opinion words have long range dependency. Their experiment results indicate that CRF does not perform well. As continuous vector space representations of words have been widely used in NLP domain, which open a new research perspective for feature extraction [15].

In this work, we aim to extract the explicit product feature (feature in brief), and define local context information and global context information as follows:

Local context information (LCI): Direct links of opinion words and noun words in a clause, e.g., co-occurrence.

Global context information (GCI): All direct and indirect links between opinion words and noun words contained in the corpus. Indirect links are links through inter-sentences. We will give an example in section 2.

The studies that follow the ideas in [3] mostly use only LCI, while GCI is not taken into account. This work combines LCI and GCI. This paper makes these contributions:

1. It shows that feature extraction can benefit from GCI.
2. It integrates LCI and GCI for extraction, which outperforms both LCI and GCI based approaches individually. This is verified through our experimental results.

After extraction, ranking features is also desirable as it helps user find important features. Feature ranking is based on feature score and frequency. The intuition is that if a feature candidate is correct and frequently mentioned in the corpus, it should be ranked high; otherwise it should be ranked low [26].

2. The proposed method

We start with the discussion of local context information (LCI) and global context information (GCI) with the following example sentences:

Example 1: "*The camera takes great pictures*."
Example 2: "*The photos are great*."
Example 3: "*This is an amazing photo*."

Each example sentence has a product feature modified by an opinion word. We can use co-occurrences between noun words and opinion words to extract the LCI links: $\langle picture, great\rangle$, $\langle photo, great\rangle$, and $\langle photo, amazing\rangle$. These direct pair-wise relations can be used to construct a bipartite graph. HITS algorithm can be applied to compute the authority scores of noun words (details in section 2.1). Meanwhile we notice that there is an indirect link between "*picture*" and "*amazing*" ($picture \rightarrow great \rightarrow photo \rightarrow amazing$). Indirect links are not only helpful for feature candidate extraction, but also for feature ranking. HITS does not consider indirect links such as this. But if we want to capture all such indirect links, we have to analyze the whole corpus to get GCI.

The SimRank algorithm proposed by Jeh and Widom [6], which measures relevance of the structural context in which objects occur, can consider indirect links in our case. Thus it can be adopted in our feature extraction task, if we construct a corpus graph by using all the direct and indirect links among opinion words and noun words. After that, we can compute the relevance of two words in an opinion word list and a noun word list using SimRank. The key difference between HITS and SimRank is that SimRank can measure all direct and indirect links in the corpus, while HITS only uses direct links. That is the reason why we apply SimRank on GCI for feature extraction. We thus designed three steps to identify and rank product features:

1. Feature Identification by LCI: This step collects all the co-occurrence pairs between noun words and opinion words in sentences, and models the problem as a bipartite graph. HITS is applied to compute the authority scores for all noun words.
2. Feature Identification by GCI: It takes all noun words and opinion words into account, and each word is a node of the global graph. If two words, which are contained in an opinion word list or a noun word list, co-occur in a sentence, there are direct links between them, and indirect links are from inter-sentences. Then a global graph is constructed. SimRank is then applied to compute the relevance of two nodes in the global graph. An example is given in Fig. 2 below in section 2.2.
3. Feature ranking: This step ranks feature candidates based on feature frequency and the combination of feature scores from the two steps above (in section 2.3). The ranking helps users find more important features.

2.1. Feature identification by LCI

The idea here is that if an adjective opinion word modifies many noun features, it is highly likely to be a good opinion word. If a feature candidate is modified by many adjective opinion words, it is likely to be a genuine feature. The problem is modeled with a bipartite graph, with noun words (feature candidates) as authorities and opinion words as hubs. The HITS algorithm is adopted to obtain the authority and hub scores.

Since our research focuses on extracting and ranking features, we assume the direction of the edge is from opinion words to feature candidates, which is illustrated in Fig. 1. If a feature candidate has a high authority score, it