Knowledge-Based Systems 56 (2014) 191–200

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Automatic construction of domain-specific sentiment lexicon based on constrained label propagation

Sheng Huang [a], Zhendong Niu [a,b,c,*], Chongyang Shi [a]

[a] School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China
[b] Beijing Engineering Research Center of Massive Language Information Processing and Cloud Computing Application, Beijing 100081, China
[c] School of Information Sciences, University of Pittsburgh, Pittsburgh, PA 15260, USA

ARTICLE INFO

Article history:
Received 26 February 2013
Received in revised form 13 September 2013
Accepted 8 November 2013
Available online 20 November 2013

Keywords:
Automatic construction
Domain-specific sentiment lexicon
Constraint propagation
Constrained label propagation
Opinion mining

ABSTRACT

Domain-specific sentiment lexicon has played an important role in most practical opinion mining systems. Due to the ubiquitous domain diversity and absence of domain-specific prior knowledge, automatic construction of domain-specific sentiment lexicon has become a challenging research topic in recent years. This paper proposes a novel automatic construction strategy of domain-specific sentiment lexicon based on constrained label propagation. The candidate sentiment terms are extracted by leveraging the chunk dependency information and prior generic lexicon. The pairwise contextual and morphological constraints are defined and extracted between sentiment terms from the domain corpus, and are exploited as prior knowledge to improve the sentiment lexicon construction. The constraint propagation is applied to spread the effect of local constraints throughout the entire collection of candidate sentiment terms. The final propagated constraints are incorporated into the label propagation for the domain-specific sentiment lexicon construction. Experimental results on real-life datasets demonstrate that our approach to constrained label propagation could dramatically improve the performance of automatic construction of domain-specific sentiment lexicon.

1. Introduction

Opinion mining, also known as sentiment analysis, is the task of detecting and extracting opinionated aspects and related user opinions in subjective texts. In recent years, the rapid growing volume of user-generated contents on the Web has greatly boosted its research activities and practical applications. Various opinion mining tasks have been conducted deeply and broadly, such as sentiment extraction and classification, sentiment summarization, sentiment spam detection, opinion search and ranking. It is increasingly becoming a practice of great value to market intelligence, information search and personal recommendation.

Since sentiment words and phrases are the basic linguistic units of expressing opinions, sentiment lexicon has played an important role in recognizing the given text's sentiment polarity. A sentiment lexicon is a list of sentiment words and phrases, which are used to indicate sentiment polarities (e.g., positive and negative). However, the sentiment lexicon is domain-dependent as users may use different sentiment words to express their opinions in different domains. This intrinsic domain-dependent characteristic makes it a tedious and laborious task to manually construct the sentiment lexicon, and a challenging problem to automatically construct for each target domain. Therefore, with the explosion of various domains corpora, how to automatically or semi-automatically construct the sentiment lexicon for the new target domain is becoming a fundamental task in opinion mining [1,2,5,13,15,19,26].

This paper focuses on the problem of automatic domain-specific sentiment lexicon construction. By definition, a domain-specific sentiment lexicon commonly contains a collection of sentiment bearing terms and phrases, and their sentiment polarities in the specific domain. The previous studies aimed at solving this problem could be mainly divided into two categories: the semantic thesaurus based sentiment lexicon construction and the corpus statistics based sentiment lexicon construction. The first category [3,5–7] utilizes the semantic relations and glosses in existing thesaurus to determine the sentiment polarities of words. The second category [1,13,15,19] exploits the statistical co-occurrence information in large domain corpus, which is based on a hypothesis that polar terms conveying the same polarities co-occur with each other. Typically, a small set of sentiment seeds are prepared for both polarities, and the polarities of candidate sentiment terms are determined based on the strength of semantic associations with the seeds [2]. However, the first category of studies ignored the domain-specific characteristic of sentiment lexicon, and could not be applied to languages without prior thesaurus knowledge.

* Corresponding author.
E-mail address: zniu@bit.edu.cn (Z. Niu).

0950-7051/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2013.11.009

The second category of studies only depended on the statistical co-occurrence in domain corpus, without consideration of the prior constraints knowledge.

The primary goal of this work is to investigate how to define and extract some prior constraints knowledge for sentiment terms, and incorporate them into a unified learning framework for the domain-specific sentiment lexicon construction. Two intuitive assumptions are fundamental in this work: (1) Each domain-specific sentiment lexicon could be divided into two parts: the domain-dependent part and the domain-independent part. The illustration is presented in Fig. 1. The shadowed overlapped area denotes the domain-independent part shared between domains A and B, and the remainder areas represent the domain-specific parts for domains A and B respectively. The domain-independent sentiment terms are usually used across multiple domains and keep consistent sentiment polarities, so the sentiment information can be propagated from the domain-independent sentiment terms to the domain-dependent sentiment terms based on their domain-specific semantic associations. (2) The sentiment terms always follow the contextual and morphological constraints throughout the domain corpus. For example, the morphological antonyms "helpful" and "helpless" always keep opposite sentiment polarities across multiple domains.

In this work, we propose an automatic construction strategy of domain-specific sentiment lexicon based on constrained label propagation. The candidate sentiment terms are extracted from domain corpus by leveraging the chunk dependency knowledge and prior generic sentiment lexicon. The pairwise constraint relations are defined and extracted between sentiment terms from domain corpus, and are exploited as prior knowledge to improve the sentiment lexicon construction. The contextual and morphological types of constraints are mainly utilized. The constraint propagation algorithm is employed to propagate the local constraints throughout the entire collection of candidate sentiment terms. The final propagated constraints are incorporated into the constrained label propagation for the sentiment lexicon construction. Since our strategy is semi-supervised, some sentiment seeds are also needed to trigger the propagation learning process. More specifically, the whole strategy can be divided into six steps: Sentiment terms extraction firstly detects and extracts candidate domain-specific sentiment terms by combining the chunk dependency parsing knowledge and prior generic sentiment lexicon, some filtering and pruning operations are performed to refine the candidate sentiment terms; Sentiment seeds extraction selects some representative domain-independent sentiment seeds from the semi-structured domain reviews, which also can be designated manually or directly borrowed from other domains when the semi-structured reviews are unavailable; Association similarity graph construction calculates the semantic associations between sentiment terms based on their distribution contexts in domain corpus; Then constraints definition and extraction try to define and extract some pairwise contextual and morphological constraints between sentiment terms to enhance their associations; And constraint propagation is applied to spread the local constraints throughout the entire collection of candidate sentiment terms; At last, the final propagated constraints are incorporated into the label propagation for the construction of domain-specific sentiment lexicon.

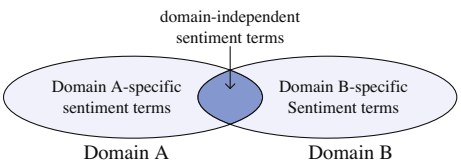

Fig. 1. The illustration of two domain-specific lexicons overlapping.

The rest of this paper is organized as follows. Section 2 surveys related work. Section 3 describes each step in our sentiment lexicon construction strategy in details. Then Section 4 presents and discusses the empirical experiments. Lastly Section 5 concludes this paper and discusses future work.

2. Related work

Many research studies have been addressed to the sentiment lexicon construction problem in recent years. Most of them utilize some paradigm words and word similarities to construct the sentiment lexicon. According to the manner of obtaining word similarities, these studies can be mainly classified into two categories of approaches: the semantic thesaurus based approaches and the domain corpus based approaches.

2.1. Semantic thesaurus based approaches

Some studies have proposed to utilize existing semantic thesauruses for sentiment lexicon construction, like WordNet, General Inquirer and Open office thesaurus for English, Hownet and CiLin for Chinese, etc. This category of studies mainly depended on the synonym and antonym relationships among sentiment terms and the grosses in thesauruses to expand the polarity lexicon from several basic sentiment seeds.

Hu and Liu [3] utilized the adjective synonym and antonym sets in WordNet to predict the semantic orientations for adjectives. Kamps et al. [4] built a lexical network by linking synonyms provided by the thesaurus, and the sentiment polarity was defined by the distance from seed words ("good" and "bad") in the network. Esuli et al. [5–7] proposed to utilize the synsets and glosses in WordNet as the sentiment units, respectively annotated their positive, neutral and negative polarity scores. The annotation process was divided into a semi-supervised learning stage and a random walking learning stage. Andrea and Fabrizio [8] further proposed the inverse model and bidirectional model of random walking algorithm, they proved its improvements over the former work in [7]. These methods commonly rely on the assumption that adjectives share the same polarities with their synonyms and opposite polarities with their antonyms.

Compared with our proposed strategy, the semantic thesaurus based approaches totally rely on prior semantic thesaurus resources without consideration for the domain-dependent characteristic of sentiment lexicon, and could not be applied to languages without such thesaurus resources. In addition, they are facing with the problem of lacking scalability. It is difficult to handle such words that are not contained in the thesaurus.

2.2. Domain corpus based approaches

The domain corpus based approaches are more widely studied in recent years. They are built on a basic assumption that polar terms conveying the same polarities co-occurred with each other in domain corpus. Contextual evidences are commonly used for the polarity assignment.

Turney [9] is one of the most primary studies that discussed learning polarities from corpus. The adjective and adverb phrases were firstly extracted as candidate sentiment terms using several pattern rules, and their polarity values were determined based on co-occurrence with two seed words ("excellent" and "poor"). The co-occurrence could be measured by the number of hits returned by a search engine. Hatzivassiloglou and McKeown [10] constructed a lexical network and determined polarity of