Neural Networks 58 (2014) 50–59

Contents lists available at ScienceDirect

Neural Networks

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

2014 Special Issue

Semi-supervised word polarity identification in resource-lean languages

Iman Dehdarbehbahani [a], Azadeh Shakery [a,b,*], Heshaam Faili [a,b]

[a] School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran

[b] School of Computer Science, Institute for Research in Fundamental Sciences (IPM), P.o. Box 19395-5746, Tehran, Iran

ARTICLE INFO

Article history:
Available online 4 June 2014

Keywords:
Sentiment lexicon
Random walk model
Semi-supervised polarity identification

ABSTRACT

Sentiment words, as fundamental constitutive parts of subjective sentences, have a substantial effect on analysis of opinions, emotions and beliefs. Most of the proposed methods for identifying the semantic orientations of words exploit rich linguistic resources such as WordNet, subjectivity corpora, or polarity tagged words. Shortage of such linguistic resources in resource-lean languages affects the performance of word polarity identification in these languages. In this paper, we present a method which exploits a language with rich subjectivity analysis resources (English) to identify the polarity of words in a resource-lean foreign language. The English WordNet and a sparse foreign WordNet infrastructure are used to create a heterogeneous, multilingual and weighted semantic network. To identify the semantic orientation of foreign words, a random walk based method is applied to the semantic network along with a set of automatically weighted English positive and negative seeds. In a post-processing phase, *synonym* and *antonym* relations in the foreign WordNet are used to filter the random walk results. Our experiments on English and Persian languages show that the proposed method can outperform state-of-the-art word polarity identification methods in both languages.

1. Introduction

Sentiment analysis or opinion mining research field deals with opinionated sentences which express people's opinions, emotions, sentiments or attitudes. Due to the rapid growth of user-generated contents on Web 2.0 sites (blogs, forums, Tweeter, and Facebook), identifying the polarity of opinionated sentences has attracted attention both in industrial and research areas. News articles, Facebook comments, product reviews and tweets are some examples of contexts in which subjective sentences are used frequently. These sentences can be informal, formal, short or long and may be used in different applications. Nevertheless, the basic idea behind identifying their polarities is identical. Automatically classifying the polarity of sentences has been studied extensively (Kim & Hovy, 2004; Pang & Lee, 2004; Taboada, Brooke, Tofiloski, Voll, & Stede, 2011; Turney, 2002).

Although the focus of natural language processing systems is moving from word-level to semantic-level and pragmatic-level techniques (Cambria & White, 2014), shortage in concept-level resources has caused the semantic-level techniques to be in their infancy in resource-lean languages. Thus, words and phrases play an important role in sentiment analysis (Liu, 2012), specially for resource-lean languages. Sentiment lexicons are composed of sentiment words, phrases and idioms which convey a positive or negative feeling or express an opinion. Each entry in the sentiment lexicon has a positive, negative, or neutral polarity. For example "good", "excellent" and "nice" have positive orientations and "bad", "obsolete" and "hate" have negative orientations. Some sentiment lexicons further assign a polarity score to each entry.

Different types of sentiment lexicon entries have been considered in the previous studies. In some research, polarity is assigned to senses of words (Esuli & Sebastiani, 2006) and in some others, polarity is assigned to expressions or idioms (Johansson & Moschitti, 2011; Takamura, Inui, & Okumura, 2007). Polarity can also be assigned to a combination of word's lemma and its attached Part of Speech (POS) tag (Hassan & Radev, 2010; Takamura, Inui, & Okumura, 2005). We refer to these types of entries as *lemmaPOS*.

The polarity of words can be identified using different approaches. Polarities can be determined manually, which is a time-consuming and laborious task. Dictionary-based and corpus-based approaches are two main approaches for identifying word polarities automatically. In corpus-based approaches, a list of words

* Corresponding author at: School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran. Tel.: +98 21 61119722.
E-mail addresses: i.dehdar@ut.ac.ir (I. Dehdarbehbahani), shakery@ut.ac.ir (A. Shakery), h.faili@ut.ac.ir (H. Faili).

http://dx.doi.org/10.1016/j.neunet.2014.05.018
0893-6080/

with known polarities and a collection of documents are used to assign polarities to other words. The growth of dictionaries (e.g., WordNet, thesaurus) has encouraged researchers to consider dictionary-based approaches for identifying word polarities. In dictionary-based methods, it is assumed that the relations between words in a dictionary convey a sentimental orientation between words. Some relations like *synonyms* and *hypernyms* convey the same orientation between two words, while *antonym* relations convey an inverse sentiment orientation. Most of the dictionary-based methods use a preliminary collection of words with known polarities as seed words. They exploit the seeds and the relations between words to identify the polarities of other words. In a basic approach, one can use WordNet and its relations in a bootstrapping manner to assign a polarity to each word. Both dictionary-based and corpus-based approaches need subjectivity analysis resources (e.g., WordNet, polarity tagged words, or opinionated corpora) to identify the word polarities. Resource-lean languages confront with a shortage in their subjectivity analysis resources, which may affect the performance of dictionary-based methods in such languages.

Recently, some research has been done which take sentiment analysis a step further from word-level to concept-level (Cambria, Olsher, & Rajagopal, 2014; Poria et al., 2013; Tsai, Wu, Tsai, & Hsu, 2013). These approaches generally utilize rich linguistic resources. ConceptNet (Speer & Havasi, 2013) as a semantic network of concepts, a biologically-inspired and psychologically-motivated emotion categorization model called Hourglass of Emotions (Cambria, Livingstone, & Hussain, 2012) and AffectiveSpace (Cambria, Hussain, Havasi, & Eckl, 2010) are samples of these resources. Although these resources provide rich information for sentiment analysis, they may not be available in resource-lean languages.

In this paper, we propose a semi-supervised method to identify the polarity of lemmaPOSs. Resources from a rich resource language (English) are incorporated to cover the resource shortage in resource-lean languages (foreign). A sparse foreign WordNet and the rich English WordNet infrastructure are exploited as sources to build a heterogeneous, multilingual, weighted and directed semantic network. Two relatedness graphs are built for the foreign and English languages. Both of the relatedness graphs are founded based on the senses and the relations between them. To connect the foreign graph to the English graph, the mappings from foreign senses to the English senses which are available in the foreign WordNet are used. For cases where the mappings are incomplete, an online translator is used to connect the two graphs. The edges between senses are weighted and directed, and they may have positive or negative weights. Our goal is to identify the polarity of *lemmaPOSs*, thus, the *lemmaPOSs* are connected to the corresponding senses in a weighted manner based on the commonness of the senses. To assign a polarity score to each foreign *LemmaPOS*, a random walk based model is applied to the semantic network along with a set of weighted English seeds in a semi-supervised manner. The random surfer assigns a polarity score to each foreign *lemmaPOS* based on the weighted seeds in the neighborhood of the foreign *lemmaPOS*. A word with different POSs may have different semantic orientations. For example the word "keen" has different semantic orientations when is used as an adjective or a noun. Thus the proposed method assigns the polarity to *lemmaPOSs*, and the senses and the relations between them are exploited to increase the accuracy of polarity identification. To improve the performance, in a post-processing phase the random walk results are purified using foreign *synonym* and *antonym* relations.

Throughout this paper, by resource-lean languages we refer to languages which do not have WordNet or have a sparse WordNet which only contains the primary relations like *synonym* and *antonym* relations. The proposed method can also be applied to resource-lean languages which do not have WordNet, in which case we should neglect the post-processing phase. In cases where the resource-lean language has a sparse WordNet, we can apply post-processing to improve the results. In both cases, we need to use a mapping between words in the resource-lean and the resource-rich languages.

Persian and English languages are chosen as foreign and rich resource languages in our empirical experiments respectively. We evaluate the performance of the proposed method for Persian and English languages, and investigate the impact of different parts of the proposed method. The experiments show that the proposed method performs well in discriminating between polar and non-polar *lemmaPOSs* and also in assigning polarity scores to polar *lemmaPOSs*, in both Persian and English languages.

The paper continues in Section 2 with reviewing some related works. In Section 3 we briefly describe our proposed method. We demonstrate the details of our experiments in Section 4 and finally conclude in Section 5.

2. Related work

Identifying semantic orientation of words has been considered in various research studies. Corpus-based (Hatzivassiloglou & McKeown, 1997; Kaji & Kitsuregawa, 2007; Turney, 2002; Velikovich, Blair-Goldensohn, Hannan, & McDonald, 2010) and dictionary-based (Esuli & Sebastiani, 2007; Kamps, Marx, Mokken, & Rijke, 2004; Rao & Ravichandran, 2009; Takamura et al., 2005) approaches are two main lines of study in this research area. Supervised approaches have also been considered by some researchers (Poria et al., 2013; Weichselbraun, Gindl, & Scharl, 2013) for this task, but these approaches require labeled data for identifying semantic orientation of words. Since these labeled data are not available in resource-lean languages, most of the researchers consider semi-supervised or unsupervised approaches in these situations.

2.1. Corpus-based approaches

Hatzivassiloglou and McKeown (1997) propose a method based on linguistic rules to identify adjective polarities. Indirect information (e.g., conjunctive between adjectives) extracted from a large corpus is used as a clue for sentiment classification of words. They show that the conjunctions between adjectives (e.g., "AND" and "OR") provide useful information about semantic orientation of adjectives.

Turney (2002) uses some linguistic rules to extract phrases from a corpus and then computes Pointwise Mutual Information (PMI) between extracted phrases and two reference words "excellent" and "poor". The search engine hit counts between the extracted phrases and the reference words are used for estimation of polarity scores.

Velikovich et al. (2010) exploit extracted n-grams from four billion Web pages as nodes to create a graph. The edges between nodes are assigned weights based on the cosine similarity between corresponding phrase context vectors. They propagate the sentiment orientations through the edges of the graph, from positive and negative seeds to other phrases.

Many research studies have been done to build sentiment lexicons for English language, but a few research studies have investigated other languages (Hassan, Abu-Jbara, Jha, & Radev, 2011; Kaji & Kitsuregawa, 2007; Kanayama & Nasukawa, 2006; Kim & Hovy, 2004).

Kaji and Kitsuregawa (2007) exploit some characteristics of Japanese language and some cues (e.g., "pros" and "cons") for extracting opinionated sentences from HTML documents. In the next step, they count the number of times a polar phrase occurred