Decision Support Systems 62 (2014) 43–53

Contents lists available at ScienceDirect

Decision Support Systems

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

Multi-lingual support for lexicon-based sentiment analysis guided by semantics

Alexander Hogenboom [a], Bas Heerschop [a], Flavius Frasincar [a,*], Uzay Kaymak [b], Franciska de Jong [a,c]

[a] Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands
[b] Eindhoven University of Technology, P.O. Box 513, NL-5600 MB Eindhoven, The Netherlands
[c] Universiteit Twente, P.O. Box 217, NL-7500 AE Enschede, The Netherlands

ARTICLE INFO

Article history:
Received 24 September 2013
Received in revised form 11 February 2014
Accepted 11 March 2014
Available online 20 March 2014

Keywords:
Multi-lingual sentiment analysis
Semantics
Lexicon
Machine translation
Map
Propagation

ABSTRACT

Many sentiment analysis methods rely on sentiment lexicons, containing words and their associated sentiment, and are tailored to one specific language. Yet, the ever-growing amount of data in different languages on the Web renders multi-lingual support increasingly important. In this paper, we assess various methods for supporting an additional target language in lexicon-based sentiment analysis. As a baseline, we automatically translate text into a reference language for which a sentiment lexicon is available, and subsequently analyze the translated text. Second, we consider mapping sentiment scores from a semantically enabled sentiment lexicon in the reference language to a new target sentiment lexicon, by traversing relations between language-specific semantic lexicons. Last, we consider creating a target sentiment lexicon by propagating sentiment of seed words in a semantic lexicon for the target language. When extending sentiment analysis from English to Dutch, mapping sentiment across languages by exploiting relations between semantic lexicons yields a significant performance improvement over the baseline of about 29% in terms of accuracy and macro-level F_1 on our data. Propagating sentiment in language-specific semantic lexicons can outperform the baseline by up to about 47%, depending on the seed set of sentiment-carrying words. This indicates that sentiment is not only linked to word meanings, but tends to have a language-specific dimension as well.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

In today's complex, globalizing markets, information monitoring tools are of paramount importance for decision makers. Such tools help decision makers in identifying issues and patterns that matter, as well as in tracking and predicting emerging events. Traditional decision support systems typically provide support for decisions by accurately deriving actionable knowledge from structured data, whereas the extraction of useful information from unstructured data like natural language text still poses important challenges [1]. Recent advances in tools for information extraction have been primarily focused on retrieving explicit pieces of information from natural language text on different levels of granularity [2]. State-of-the-art information monitoring and extraction tools enable us to identify entities like companies, products, or brands in text, and to subsequently extract more complex concepts, such as events in which these entities play various roles [3]. Recent research endeavors additionally explore how to perform such information extraction tasks on a multitude of heterogeneous sources in an ever-changing environment [4–6].

However, latent pieces of information can be extracted from natural language text as well. For instance, recent work has made it possible to detect the distinct topics that people discuss in their (on-line) conversations [7,8]. Yet, for many application scenarios, it is not so much the entities, events, or topics that people discuss per se, but rather people's sentiment with respect to these subjects that provides decision makers with valuable information. This is reflected by the recent surge in research interest in sentiment analysis for decision support [1,9–11].

Sentiment analysis techniques can support decision making in a multitude of scenarios. For instance, sentiment analysis can help organizations pinpoint the effect of specific issues on customer perceptions, thus helping these organizations respond with appropriate marketing and public relations strategies [12]. Furthermore, consumer sentiment has been demonstrated to have a significant impact on stock ratings [13,14] and sales [15,16]. Thus, accurate sentiment analysis methods are crucial for supporting decision making in these fields. Additionally, tracking of stakeholders' sentiment is important for decision making in economic systems [17], financial markets [18], politics [19], organizations [20], and reputation management [21].

Real-world decision support systems typically consist of four logical components, i.e., a Knowledge Management System (KMS), a Model Management System (MMS), a Database Management System (DMS), and a User Interface System (UIS) [22]. Each of these logical components

* Corresponding author. Tel.: +31 10 408 1340; fax: +31 10 408 9162.
E-mail addresses: hogenboom@ese.eur.nl (A. Hogenboom), basheerschop@gmail.com (B. Heerschop), frasincar@ese.eur.nl (F. Frasincar), u.kaymak@ieee.org (U. Kaymak), f.m.g.dejong@utwente.nl (F. de Jong).

http://dx.doi.org/10.1016/j.dss.2014.03.004
0167-9236/© 2014 Elsevier B.V. All rights reserved.

is used to monitor and regulate the flow of crucial information in order to support decision making in an organization. Data managed by the DMS can be transformed into actionable knowledge in the KMS, with the MMS controlling how the obtained knowledge is used in models in order to support decision making, and the UIS taking care of the interaction with the end user of the system. In order to utilize sentiment-based information in decision support systems, the DMS should be enriched with (user-generated) sentiment-carrying content that has been crawled from the Web. Furthermore, the KMS should be able to represent the indicators of identified sentiment with respect to a topic of interest. Additionally, the MMS should allow for the incorporation of sentiment-based information in the decision making process. Last, the UIS should provide dashboards with relevant information that enables decision makers to act upon arising issues in a timely manner.

One of the key open issues that must be resolved in order to be able to exploit the full potential of sentiment analysis in real-life decision support systems is that these systems must be able to deal with textual data in various languages [1]. Such data is available in vast amounts, as recent developments on the Web enable users to produce an ever-growing amount of virtual utterances of opinions or sentiment through, e.g., messages on Twitter, blogs, or reviews, in any language of their preference.

The analysis of sentiment in the overwhelming amount of available multi-lingual textual data is challenging at best. This challenge can be addressed by means of automated sentiment analysis techniques, focusing on determining the polarity of natural language text. Typical approaches involve scanning a text for cues signaling its polarity, e.g., (parts of) words or other (latent) features of natural language text. Lexicon-based sentiment analysis methods have gained (renewed) attention in recent work [23–29], not in the least because their performance has been shown to be robust across domains and texts [30]. Such methods essentially rely on lexical resources containing words and their associated sentiment, i.e., sentiment lexicons, and their nature allows for intuitive ways of accounting for structural or semantic aspects of text in sentiment analysis [26,31].

Many existing lexicon-based sentiment analysis approaches are tailored to one specific language — typically English. However, in order for automated sentiment analysis to be useful for decision makers in today's complex, globalizing markets, automated sentiment analysis tools need to be able to support multiple languages rather than English only. Therefore, we explore how we can analyze sentiment in another language – i.e., Dutch – for which we have nothing more but some lexical and syntactical parsing tools, a semantic lexical resource, and a handful of positive and negative sample words.

A good starting point is SentiWordNet [32,33], as recent research has proven this large (semantic) sentiment lexicon for English, generated by means of machine learning techniques, to be rather effective when used for analyzing sentiment in texts published in our reference language, i.e., English [34]. As a first step, one could consider translating texts from a target language, i.e., Dutch, to our reference language, i.e., English, in order to be able to subsequently utilize the well-established SentiWordNet sentiment lexicon for the reference language in the sentiment analysis process.

However, as subjectivity is associated with word meanings rather than words [35], the literal translation of texts to a reference language in order to benefit from the available sentiment lexicon for the reference language may be suboptimal in automated sentiment analysis of texts in another language. As an alternative, we therefore propose to map the sentiment from the reference sentiment lexicon to a sentiment lexicon for the target language, by means of traversing relations between large language-specific semantic lexical resources, thus accounting for word meanings rather than lexical representations. Additionally, we consider an approach that involves propagating sentiment from a seed set of words in a language-specific semantic lexical resource for each considered language separately, in order to generate language-specific sentiment lexicons which can subsequently be used in language-specific sentiment analysis methods.

The main contribution of our work lies in our novel sentiment mapping method, which exploits relations between language-specific semantic lexicons in order to construct a sentiment lexicon for a target language. We compare the effectiveness of this method with that of an existing machine-translation approach and a method that focuses on semantic relations within, rather than across languages. We thus aim to provide insight in the importance of semantics for multi-lingual sentiment analysis.

The remainder of the paper is organized as follows. In Section 2, we discuss related work on (multi-lingual) sentiment analysis and the semantic lexicons that may be exploited in this process. We then elaborate on our framework for assessing our considered methods for dealing with another language in sentiment analysis in Section 3. Our findings are discussed in Section 4. We conclude and provide directions for future work in Section 5.

2. Related work

Today's abundance of user-generated content has resulted in a surge of research interest in systems that are able to deal with opinions and sentiment, as explicit information on user opinions is often hard to find, confusing, or overwhelming [36]. Many language-specific sentiment analysis approaches exist, whereas the exploration of how to support multiple languages when analyzing sentiment has only just begun.

2.1. Sentiment analysis

The roots of sentiment analysis are in fields like natural language processing, computational linguistics, and text mining. The main objective of most sentiment analysis approaches is to extract subjective information from natural language text. Most work focuses on determining the overall polarity of words, sentences, text segments, or documents [36]. This task is commonly approached as a binary classification problem, in which a text is to be classified as either positive or negative. However, this task may be approached as a ternary classification problem as well, by introducing a third class of neutral documents. An alternative to such sentiment classification approaches is the determination of a degree of positivity or negativity of natural language text in order to produce, e.g., rankings of positive and negative documents [37,38].

Many state-of-the-art approaches to sentiment classification tasks rely on machine learning techniques [36,39]. On the other hand, some approaches exploit (generic) sentiment lexicons when determining the subjectivity or polarity of natural language text. Both approaches may be combined in hybrid methods as well [29].

In machine learning sentiment analysis methods, natural language text is typically modeled by means of a bag-of-words vector representation, denoting an unordered collection of words occurring in this text. In order to be able to, e.g., distinguish pieces of text from one another in terms of their associated polarity class, machine learning methods typically aim to find and exploit patterns in the vector representations of these texts. In such vector representations, a binary encoding scheme, indicating the presence of specific words, has proven to be effective [39] as well as to outperform frequency-based encoding [40]. Vectors may also contain features other than words, e.g., parts of words, word groups, or features representing semantic distinctions between words [41]. Features represented in vectors may be weighted as well [42].

Lexicon-based methods account for the semantic orientation of individual words in a text by matching these words with a list of words and their associated sentiment scores, i.e., a sentiment lexicon. The text's overall semantic orientation is then determined by aggregating (e.g., summing) the individual word scores, as retrieved from the sentiment lexicon. Hybrid approaches may realize the aggregation through a machine learning process as well [29]. In this sentiment scoring process, other aspects of content may be taken into account as well, such as

Download English Version:

https://daneshyari.com/en/article/554730

Download Persian Version:

https://daneshyari.com/article/554730

Daneshyari.com