Data & Knowledge Engineering 94 (2014) 135–147

Contents lists available at ScienceDirect

Data & Knowledge Engineering

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

Editorial

Rhetorical Structure Theory for polarity estimation: An experimental study

José M. Chenlo [a,*], Alexander Hogenboom [b], David E. Losada [a]

[a] Centro de Investigación en Tecnoloxías da Información (CITIUS), University of Santiago de Compostela, Spain
[b] Econometric Institute, Erasmus University Rotterdam, The Netherlands

ARTICLE INFO

Article history:
Received 30 October 2013
Accepted 7 July 2014
Available online 18 July 2014

Keywords:
Information Retrieval
Text mining
Sentiment analysis
Natural Language Processing
Rhetorical Structure Theory

ABSTRACT

Sentiment analysis tools often rely on counts of sentiment-carrying words, ignoring structural aspects of content. Natural Language Processing has been fruitfully exploited in text mining, but advanced discourse processing is still nonpervasive for mining opinions. Some studies, however, extracted opinions based on the discursive role of text segments. The merits of such computationally intensive analyses have thus far been assessed in very specific, small-scale scenarios. In this paper, we investigate the usefulness of Rhetorical Structure Theory in various sentiment analysis tasks on different types of information sources. First, we demonstrate how to perform a large-scale ranking of individual blog posts in terms of their overall polarity, by exploiting the rhetorical structure of a few key evaluative sentences. In order to further validate our findings, we additionally explore the potential of Rhetorical Structure Theory in sentence-level polarity classification of news and product reviews. Our most valuable polarity classification features turn out to capture the way in which polar terms are used, rather than the sentiment-carrying words per se.

1. Introduction

Natural Language Processing (NLP) has become of vital importance for current information systems [1]. Recent advances in NLP permit to distill actionable knowledge from massive amounts of Web content: by detecting important events in news [2]; by discovering topics and viewpoints from social media [3–5]; or by associating news messages and social media posts with their potential effects on stock prices [6,7] or sales [8].

Sentiment analysis (SA) – also known as opinion mining – is an active and influential research area concerned with automatically extracting subjectivity from natural language text [9–12]. It deals with tasks such as classifying the polarity of documents as positive or negative, or ranking documents in terms of their associated degree of positivity or negativity with respect to a topic of interest. With one fifth of all tweets [13], one third of all blog posts [14], and the vast majority of reviews discussing products or brands, SA is crucial for revealing traces of people's sentiment from ubiquitous user-generated content.

Commercial sentiment analysis systems mostly rely on simple occurrences of sentiment-carrying words in texts [12]. Yet, word frequencies alone are insufficient for mining sentiments [12,15–17]. Accounting for the way in which words are used is essential for text understanding. In this light, one of the key open research issues refers to the role of textual structure [12]. Structural aspects seem to be valuable for various text mining tasks [18–21], including SA [15,16,22–24], but this requires further study.

An increasingly popular way of accounting for structural aspects of opinionated text is to analyse the rhetorical organization of documents [16,24–28]. One way of accomplishing this is by applying the Rhetorical Structure Theory (RST) [29]. As a leading descriptive framework for text, RST identifies the rhetorical roles (e.g., explanation, contrast) of text segments. This is useful for SA, as

* Corresponding author. Tel.: +34 881 816 396; fax: +34 881 816 405.
E-mail addresses: josemanuel.gonzalez@usc.es (J.M. Chenlo), hogenboom@ese.eur.nl (A. Hogenboom), david.losada@usc.es (D.E. Losada).

http://dx.doi.org/10.1016/j.datak.2014.07.009
0169-023X/

sentiment-carrying words in, e.g., an explanation segment may contribute differently to the overall sentiment than those in a contrasting segment. But RST for SA has been mostly evaluated in small-scale document classification studies [16,24]. In such constrained settings, RST significantly contributed to determine polarity, at a cost of computational complexity.

In this paper, we thoroughly study RST for SA and experiment with various opinion repositories – which vary in size and source, i.e., blogs, news, and product reviews – and different polarity analysis tasks of varying granularity, i.e., document polarity ranking and sentence-level polarity classification. To the best of our knowledge, this is the first study of this kind.

In the first part of this paper, we quantify the advantage of exploiting RST for blog post polarity ranking and we identify the rhetorical relations that help to understand the sentiment conveyed in blogs. For efficiency reasons, we build upon recent advances in extracting key opinionated sentences from blog posts [30] and analyse the discourse only for selected passages. The evaluation of RST is therefore indirect, as we test how the rhetorical analysis of selected sentences helps to estimate the polarity of documents as a whole.

In the second part of this paper we focus on a fine-grained task – sentence-level polarity classification – and perform a more direct evaluation of the merits of RST. We study the ability of RST to reveal positively or negatively-oriented sentences within news articles or product reviews. Sentence-level polarity classification goes beyond document-level sentiment classification as it moves closer to opinion targets and sentiments on the targets. This facilitates opinion extraction from text that may only contain a few sentences that discuss the topic of interest [10].

The remainder of this paper is organised as follows. First, in Section 2, we discuss related work on existing SA approaches and review how structural aspects of content are typically involved in such methods. In Section 3, we describe our novel method of document-level polarity estimation that works at large scale. Section 4 focuses on sentence-level polarity classification guided by RST. Last, in Section 5, we present our conclusions and suggest directions for future research.

2. Related work

Explicit information on user opinions is often hard to find, confusing, or overwhelming [9]. The abundance of user-generated content on the Web has led to a surge of research interest in systems that automatically mine opinions and sentiment. Many of such SA systems exist, but the exploration of how to account for structural aspects of content when analysing sentiment has only just begun.

2.1. Sentiment analysis

SA tools often apply Computational Linguistics and text mining technology. Typical tasks include distinguishing subjective text segments from objective ones, as well as determining the polarity of words, sentences, or documents [9]. The latter is often approached as a two-class categorisation problem of distinguishing positive from negative text or, occasionally, as a three-class problem, in which an additional class of neutral text is considered. An alternative to polarity classification is to determine a degree of positivity or negativity of natural language text and produce, e.g., rankings of positively and negatively oriented documents.

The state-of-the-art in automated SA has been reviewed extensively [9–12]. Existing methods range from Machine Learning methods, exploiting patterns in vectorial representations of text, to lexicon-based methods, accounting for the semantic orientation of individual words (e.g., using sentiment lexicons). Many hybrid approaches exist as well.

Large-scale SA tasks typically pose unique challenges, not just in extracting sentiment from a large set of documents, but also in identifying on-topic (fragments of) documents. Numerous studies have been conducted on how to mine opinions from large-scale repositories like the blogosphere. Given a topic of interest, the search for relevant and subjective documents (regardless of their polarity) has been studied by different scientists [31–33]; and Chenlo and Losada proposed effective and efficient methods of finding opinionated segments in blog posts [30]. These methods permit to represent the overall opinion of a blog post with a limited number of sentences that are selected by combining three types of sentence-level evidence: topicality, polarity, and location.

Although relying on counts of sentiment-carrying words is still predominant [12], other aspects of content are promising. Early work with movie reviews [22,34] considered the absolute position of text segments and found that the last sentences of a document could be indicative of the overall polarity. Positional information has proven to be useful in large-scale SA tasks as well. For example, the proximity of query terms to subjective sentences in a document was used by Santos et al. to detect on-topic opinions [32]. Similarly, Gerani et al. defined a proximity-based propagation method to calculate the aggregated opinion at the position of each query term in a document [31].

A broad array of studies employed linguistic mechanisms to extract structure. For instance, Devitt and Ahmad estimated sentiment by analysing the semantic cohesion of a text [23], with limited success. More successful attempts [16,24,26–28] selected important opinion extracts from the text's rhetorical structure — obtained by, e.g., RST.

2.2. Rhetorical Structure Theory

The structure of natural language can be characterised by the rhetorical relations that hold between parts of the text. Such relations (e.g., explanations or contrasts) are important for text understanding, because they give information about how the textual segments are related to one another to form a coherent discourse. Discourse analysis is concerned with how meaning is built up in the larger communicative process. Such an analysis can be applied on different levels of abstraction, i.e., within a sentence, within a paragraph, or within a document or conversation. The premise is that each part of a text has a specific role in conveying the overall message.