Concentration Areas of Sentiment Lexica in the Word Embedding Space

Elena Razova, Vyatka State University, Kirov, Russian Federation
Evgeny Kotelnikov, Vyatka State University, Kirov, Russian Federation

ABSTRACT

Sentiment lexicons play an important role in opinion mining systems and cognitive linguistics. Previous work aimed mostly at creating sentiment lexicons, but not thorough research into their fundamental properties. In this paper the arrangement of sentiment lexica in the multidimensional space of distributed word representations is studied. A hypothesis on the existence of sentiment lexica concentration areas is introduced and it is tested on the basis of the joint analysis of the distribution of sentiment words and general lexica. The results of the test allow to confirm the proposed hypothesis and discover the words which more than 80% of the sentiment lexica is concentrated around.

KEYWORDS

Cognitive Linguistics, Distributed Word Representations, Multidimensional Space, Opinion Mining, Sentiment Lexicons, t-SNE, Word Embeddings, Word2vec

INTRODUCTION

Over the last 10-15 years, sentiment analysis and opinion mining have become one of the most rapidly growing fields in computational linguistics (Hemmatian & Sohrabi, 2017; Sun et al., 2017; Yadollahi et al., 2017). Sentiment is the expression of subjectivity as either a positive or negative opinion (Taboada, 2016). In view of the emergence of Web 2.0, users can generate new content for websites, e.g. on blogs and forums, in the comments sections of the news website, in the social networks and so on. This content often consists of users’ opinions on different issues (products and services, political and social spheres, etc.). Extraction and investigation of such opinions could bring valuable information to owners of websites and other stakeholders. Sentiment analysis and opinion mining software tools are used in social media analysis, political and marketing research, human-computer interfaces, etc. (Liu, 2015). Also the elicitation of sentiment and opinions from texts can be useful for related fields of study such as argumentation mining (Peldszus & Stede, 2013), stance detection (Sridhar et al., 2015) and emotion recognition (Shaheen et al., 2014).

There are two main approaches to sentiment analysis and opinion mining (Sun et al., 2017): machine learning and lexicon-based approach. In the machine learning there is a need for text corpora that are labelled by sentiment (Wang et al., 2016a). On the basis of such corpora some classifier is trained. There are many types of classifiers appropriated for this task, for example, support vector

DOI: 10.4018/IJCINI.2019040104

Copyright © 2019, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
machines or deep neural networks (Wang et al., 2016b). In the lexicon-based approach the classifier is built with the use of special sentiment lexicons containing words with the highest emotional coloring (Taboada et al., 2011).

These sentiment lexicons are the key component of most opinion mining systems. Such lexicons can be created manually or automatically. Lexicons formed with the help of manual annotation are of higher quality than automatic ones, but developing them is very labor-intensive (Kotelnikov et al., 2016).

Although there is a large number of works dedicated to creating and using sentiment lexicons (Hamilton et al., 2016; Vo & Zhang, 2016; Wang & Xia, 2017), the fundamental properties of sentiment lexica have not been studied much. Recently, new neural models for distributed word representations (word embeddings) have been proposed. They show high efficiency in solving many problems of natural language processing (Mikolov et al., 2013a; Pennington et al., 2014). Knowledge of semantic relations between words in such models can help make progress in understanding the fundamental properties of sentiment lexica and to develop new algorithms for creating such lexicons.

This paper studies the arrangement of sentiment lexica in the multidimensional space of distributed word representations. We propose a hypothesis about the existence of sentiment lexica concentration areas. The hypothesis is tested on the basis of the joint analysis of the distribution of sentiment words and general lexica for the English and the Russian languages. The results of the test allow us to confirm the proposed hypothesis and discover the words which more than 80% of the sentiment lexica is concentrated around.

**RELATED WORK**

In papers dedicated to the joint application of sentiment lexicons and distributed word representations, there are two main directions: creating dictionaries based on existing models (Blinov & Kotelnikov, 2014; Garten et al., 2016; Ito et al., 2017) and building models taking into account the characteristics of sentiment lexica (Hamilton et al., 2016; Tang et al., 2014; Vo & Zhang, 2016; Wang & Xia, 2017).

Garten et al. (2016) form a general concept for the initial full moral dictionary or several seed words belonging to it by averaging vectors of distributed representations. The resulting concept can be further used to detect moral rhetoric or to expand the initial dictionary. Ito et al. (2017) build a sentiment lexicon specialized for analyzing financial policies using distributed representations of words obtained by LINE (Large-scale Information Network Embedding). The authors use bootstrapping based on the average similarity between all seed words and candidate words. Blinov and Kotelnikov (2014) build sentiment and aspect term lexicons for aspect-based sentiment analysis. For this purpose, a small number of seed words is expanded to include new words on the basis of cosine similarity in the vector space of distributed representations.

Tang et al. (2014) develop a neural network to learn sentiment-specific phrase embedding based on a corpus of tweets with emoticons. The resulting phrase embedding helps to form a sentiment lexicon by means of a small number of seed words. In (Vo & Zhang, 2016), a simple neural network is taught on a corpus of tweets with emoticons. This teaching resulted in matching each word with a two-dimensional vector, whose components denote the positive and negative coloring of a word. Hamilton et al. (Hamilton et al., 2016) build a lexical graph based on word embeddings learned on domain-specific corpora. A sentiment lexicon is formed with the help of the propagation of sentiment labels from seed words over this graph using a random walk method. In (Wang & Xia, 2017), a neural network learns sentiment-aware word embedding by using both document-level sentiment labels and word-level sentiment annotations. The sentiment lexicon is based on a classifier, taught on a small vocabulary of sentiment lexica, represented by their word embedding vectors.

In all the above-mentioned works, sentiment lexicons are created without thorough research on the space of distributed word representations. Keshi et al. (2017) associate feature words (concepts)
Text Semantic Mining Model Based on the Algebra of Human Concept Learning
Jun Zhang, Xiangfeng Luo, Xiang He and Chuanliang Cai (2013). *Cognitive Informatics for Revealing Human Cognition: Knowledge Manipulations in Natural Intelligence* (pp. 221-236).
www.igi-global.com/chapter/text-semantic-mining-model-based/72292?camid=4v1a