**Word Embeddings for Sentiment Analysis: A Comprehensive Empirical Survey**

1. Introduction

The scope of this survey work is to examine the quality of Glove andWord2vec word embeddings on word analogy task as well as on four sentiment analysis tasks: sentiment analysis of tweets, song lyrics, movie reviews, and product (phone) reviews.

2. Experimental Design

2.1. Research Questions

The purpose of this study is to examine the quality of word embeddings, observe how training method and corpus attributes influence their it, and assess their performance when used as classification features on sentiment analysis tasks.

2.2. Text Preprocessing and Training

Throughout this paper, we use the term “tokens” to denote the total number of words inside a text collection that repeat themselves in it. The total number of unique words (vocabulary size) of the collection is simply called words. In “Size” column we show the number of tokens in training corpora after preprocessing but before applying the training methods. “Dimensions” on the other hand, presents the number of float values each word of the resulting model is associated with. For consistency with the pretrained models, we trained our corpora with 300 dimensions.

2.3. Contents of Text Sets

3. Sentiment Analysis Tasks and Datasets

3.1. Sentiment Analysis of Tweets

3.2. Sentiment Analysis of Song Lyrics

Studies using lyrics are more recent and fall into two categories: lexicon-based and corpus-based.

3.3. Sentiment Analysis of Movie and Phone and Reviews

Item review sentiment analysis is about training intelligent models that are able to predict user satisfaction level from various kinds of products, based on the textual and other descriptions of those products provided by previous users.

4. Linguistic Regularity Benchmarks

One way for evaluating the quality of trained embedding models is to observe the semantic and syntactic regularities they manifest.

5. Role of Corpus Characteristics

5.1. Scores of all Models

5.2. Scores of Misch Corpora Size Cuts

Here we try to observe any role that training corpus size and training method might have on the performance of word embeddings.

5.3. Scores of Domain-specific Size Cuts

Same as in the previous subsection, in this set of experiments we also try to observe the role of model size on the performance of word embeddings.

5.4. Scores of Fixed-size Domain-specific Corpora

6. Improving Quality of Word Embeddings

The primary motive for developing distributed word representations was the need to fight the curse of dimensionality

and data sparsity problems of the classical bag-of-words representation. Word embedding training methods do however exhibit certain deficiencies that limit their performance in various applications like sentiment analysis and more. First of all, Glove andWord2vec rely on co-occurrence statistics and window-based context for training and no use of word meaning is performed. As each word is represented by a single and unique vector, they do not count for homonymy (identical spelling but different meanings) or polysemy (association of one word with two or more distinct meanings). Furthermore, in subjective and emotional texts (e.g., tweets or song lyrics) it is frequent to find words with opposite polarity appearing close to each other (same context window).

6.1. Enhancing quality with lexicons

Linguistic lexicons have been widely used in text classification tasks as a rich source of features. WordNet is one of the most popular in the literature, rich with synonymy, hypernymy, paraphrase and other types of relations between words.

6.2. Creating better embedding methods

A different direction for obtaining better word embeddings is to improve training methods in the first place.

7. Discussion

In this paper, we examined the quality of word embeddings when used in word analogy tasks, as well as their performance on sentiment analysis of tweets, song lyrics, movie reviews, and phone reviews. We were particularly interested to observe the role of factors like training method, training corpus size and thematic relevance on quality of the produced word embeddings. We also exercised the semantic and syntactic quality of models trained from different corpora with word analogy questions.