Short Text Similarity Understanding with Word Embeddings

[1]Rutuja Subhash Gadekar, [2]Prof. Bhagwan Kurhe
[1] M.E. Student, SPCOE, Otur, Pune
[2] Assistant Professor, SPCOE, Otur, Pune

Abstract - Short text messages such as tweets are very noisy and sparse in their use of vocabulary. Traditional textual representations, such as tf-idf, have difficulty grasping the semantic meaning of such texts, which is important in applications such as event detection, opinion mining, news recommendation, etc. We constructed a method based on semantic word embeddings and frequency information to arrive at low-dimensional representations for short texts designed to capture semantic similarity. For this purpose, we designed a weight-based model and a learning procedure based on a novel median-based loss function. This paper discusses the details of our model and the optimization methods, together with the experimental results on both Wikipedia and Twitter data. We find that our method outperforms the baseline approaches in the experiments, and that it generalizes well on different word embeddings without retraining. Our method is therefore capable of retaining most of the semantic information in the text and is applicable out-of-the-box.

Keyword - Short Text Similarity; Word Embeddings

1. INTRODUCTION

Short pieces of texts reach us every day through the use of social media such as Twitter, newspaper headlines, and texting. Especially on social media, millions of such short texts are sent every day, and it quickly becomes a daunting task to find similar messages among them, which is at the core of applications such as event detection (De Boom et al. (2015b)), news recommendation (Jonnalagedda and Gauch (2013)), etc.

In this paper we address the issue of finding an effective vector representation for a very short text fragment. By effective we mean that the representation should grasp most of the semantic information in that fragment. For this we use semantic word embeddings to represent individual words, and we learn how to weigh every word in the text through the use of tf-idf (term frequency - inverse document frequency) information to arrive at an overall representation of the fragment.

These representations will be evaluated through a semantic similarity task. It is therefore important to point out that textual similarity can be achieved on different levels. At the most strict level, the similarity measure between two texts is often defined as being (near) paraphrases. In a more relaxed setting one is interested in topic- and subject-related texts. For example, if a sentence is about the release of a new Star Wars episode and another about Darth Vader, they will be dissimilar in the most strict sense, although they share the same underlying subject. In this paper we focus on the broader concept of topic-based semantic similarity, as this is often applicable in the already mentioned use cases of event detection and recommendation.

Our main contributions are threefold. First, we construct a technique to calculate effective text representations by weighing word embeddings, for both fixed- and variable-length texts. Second, we devise a novel median-based loss function to be used in the context of minibatch learning to mitigate the negative effect of outliers. Finally we create a dataset of semantically related and non-related pairs of text from both Wikipedia and Twitter, on which the proposed techniques are evaluated. We will show that our technique outperforms most of the baselines in a semantic similarity task.

We will also demonstrate that our technique is independent of the word embeddings being used, so that the technique is directly applicable and thus does not require additional model training when used in different contexts, in contrast to most state-of-the art techniques.

2. MOTIVATION

As [9] introduces, supposing that a practical and valid method of calculating the semantic difference between two short texts exists, there are many applications in Natural Language Processing (NLP) that can take advantage of it. For example, in the field of information retrieval and image retrieval from the Web, one of the best techniques for improving retrieval effectiveness is by using semantic similarity.

The use of text similarity is also useful for boosting accuracy results in relevance feedback and text categorization as for methods for automatic evaluation of machine translation, evaluation of text coherence [1], word sense disambiguation, formatted documents classification and text summarization. Also, it has been proved that for data sharing systems such as federated databases, message passing or data integration systems, web services, data management systems, etc., lexical and syntactical differences between shared variables can be solved by using semantic text similarity.

Semantic text similarity can also be used to build a text similarity join operator, that can be used to join two relations if their join attributes are textually similar to each other, which can be useful in several domains, such as integration of data from heterogeneous resources, mining of data, cleansing of data, etc. [3]

3. OBJECTIVES

The objective of this thesis is to determine and prove whether a system using word embeddings generated with GloVe can perform better than state-of-the-art systems that use the collection of models Word2Vec to build the word vector representations for their final use in the field of text similarity. We compare both methods (GloVe and Word2Vec) in several ways in order to determine which aspects of the word embeddings are different for the task of semantic text similarity. After analyzing the results, we also aim to use the currently generated word embeddings with GloVe in several different ways to improve the performance of our model.

4. RELATED WORK

In this section we discuss previous work related to the different aspects of our method.

Distributional semantics.

Distributional semantic approaches are based on the intuition that words appearing in similar contexts tend to have similar meanings. The Latent Semantic Analysis algorithm (LSA) [3] incorporates this intuition by building a word-document co-occurrence matrix and performing singular value decomposition (SVD) on it to get a lower-dimensional representation. Words are represented as vectors in this lower dimensional space. The distance between these word vectors (measured, e.g., with the cosine function) can be used as a proxy for semantic similarity. The full co-occurrence matrix, however, can become quite substantial for a large corpus, in which case the SVD becomes memory-intensive and computationally expensive.

Text-level semantics with external knowledge.

Corpus methods are combined with WordNet-based measures in [13]. In [13] an IDF-weighted alignment approach, based both on WordNet-based and corpus-based similarities, is proposed. Texts are parsed and only similarities within identical part-of-speech categories are considered. Finally, a single score is calculated as an average over the maximum similarities. In a WordNet similarity measure is combined with word order scores. In neither approaches any machine learning step is applied.

SemEval STS.

Recently, the SemEval Semantic Text Similarity (STS) task [1] and SemEval STS task [2] were organised. A full description of the work of all participating teams (over 30 in both years) is beyond the scope of this section. We discuss the approaches of the best-scoring teams.The best-scoring teams in both calculate a large number of features based on a wide variety of methods. Additionally, handcrafted rules are applied that deal with currency values, negation, compounds, number overlap and with literal matching [6].

5. Existing System

Existing semantic visualization models are not designed for short texts. For example, PLSV represents documents as bags of words, and topic distributions are inferred from word co-occurrences in documents. This assumes sufficiency in word co-occurrences to discover meaningful topics. This may be valid for regular-length documents, but not for short texts, due to the extreme sparsity of words in such documents. Methods based on tf-idf vectors, such as SSE would also suffer, because tf-idf vectors are not efficient for short text analysis. Many words appear only once in a short document, and may appear in only a few documents. Consequently tf and idf are not very distinguishable in short texts.

The Existing system is a generalized framework to understand short texts effectively and efficiently. More specifically, it divide the task of short text understanding into three subtasks: text segmentation, type detection, and concept labeling. It formulate text segmentation as a weighted Maximal Clique problem, and propose a randomized approximation algorithm to maintain accuracy and improve efficiency at the same time. It introduce a Chain Model and a Pairwise Model which combine lexical and semantic features to conduct type detection. They achieve better accuracy than traditional POS taggers on the labeled benchmark. It employ a Weighted Vote algorithm to determine the most appropriate semantics for an instance when ambiguity is detected. The experimental results demonstrate that framework outperforms existing state-of-the-art approaches in the field of short text understanding. It unable to analyze and incorporate the impact of spatial-temporal features into framework for short text understanding.

6. PROPOSED SYSTEM:

Effective learning of general word semantic relations is now feasible and practical with recent developments in neural network techniques, which have contributed improvements in many tasks in Information Retrieval (IR) and Natural Language Processing (NLP). Specifically, neural network language models, e.g., Continuous Bag-of-Words (CBOW), Continuous Skip-gram model, and Glove model], learn word embeddings (or word vectors) with the aim of fully retaining the contextual information for each word, including both semantic and syntactic relations. Such general word semantic relations can be efficiently learned from a very large text corpus, in any language. In fact, there are many pre-trained word embeddings learned from resources like Wikipedia, Twitter, and Freebase, publicly available on the Web. Because of its good performance, in this paper, we propose to extend the DMM model for topic modeling over short texts by addressing its two limitations.

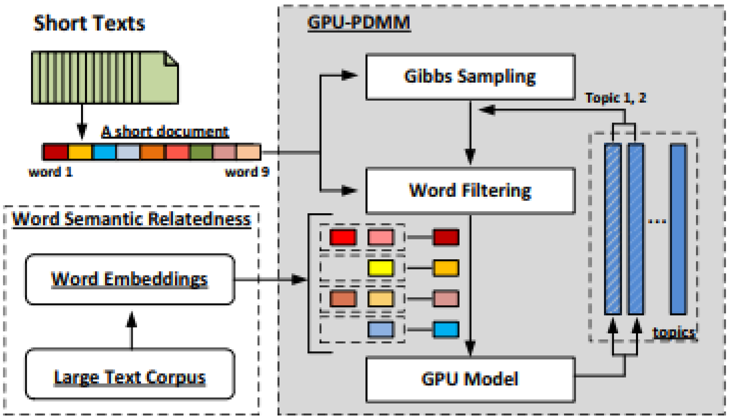

Fig.1: Proposed System Architecture

7. CONCLUSIONS AND FUTURE WORK

We have described a generic and flexible method for semantic matching of short texts, which leverages word embeddings of different dimensionality, obtained by different algorithms and from different sources. The method makes no use of external sources of structured semantic knowledge nor of linguistic tools, such as parsers. Instead it uses a word alignment method, and a saliency weighted semantic graph, to go from word-level to text-level semantics. We compute features from the word alignment method and from the means of word embeddings, to train a final classifier that predicts a semantic similarity score

We demonstrate on a large publicly available evaluation set that our generic, semantics-only method of computing semantic similarity between short texts outperforms all baseline approaches working under the same conditions, and that it exceeds all approaches using external sources of structured semantic knowledge that have been evaluated in this dataset, to our knowledge

An important implication of our results is that distributional semantics has come to a level where it can be employed by itself in a generic approach for producing features that can be used to yield state-of-the-art performance on the short text similarity task, even if no manually tuned features are added that optimise for a specific test set or domain. Furthermore, the word embeddings, when employed as proposed above, substitute external semantic knowledge and make human "feature engineering" unnecessary. As our method does not depend on NLP tools, it can be applied to domains and languages for which these are sparse

It is interesting to see how other fields of research that deal with large corpora of unstructured text can benefit. For example, in automatically created probabilistic knowledge bases (e.g., [6]) triples are extracted from an input corpus and have a confidence score associated with them based on the number of sentences in the corpus describing the relation in the triple. Short text similarity can be used to improve this confidence score.

6. REFERENCES

[1] C. Quirk, C. Brockett, and W. B. Dolan. Monolingual machine translation for paraphrase generation. In EMNLP 2004.

[2] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In ICML, 2008.

[3] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6):391–407, 1990.

[5] Y. Bengio, H. Schwenk, J.-S. Senécal, F. Morin, and J.-L. Gauvain. Neural probabilistic language models. In Innovations in Machine Learning. Springer, 2006.

[6] B. Hu, Z. Lu, H. Li, and Q. Chen. Convolutional neural network architectures for matching natural language sentences. In NIPS, 2014.

[7] A. Islam and D. Inkpen. Semantic text similarity using corpus-based word similarity and string similarity. TKDD, 2008.

[8] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. In Proceedings of the 13st International Conference on Machine Learning, 2014.

[9] Y. Kim. Convolutional neural networks for sentence classification. In EMNLP 2014, 2014.

[10] P. Shrestha. Corpus-based methods for short text similarity. Rencontre des Étudiants Chercheurs en Informatique pour le Traitement automatique des Langues, 2011.

[11] S. Fernando and M. Stevenson. A semantic similarity approach to paraphrase detection. CLUK 2008, 2008.

[12] R. Ferreira, R. D. Lins, F. Freitas, S. J. Simske, and M. Riss. A new sentence similarity assessment measure based on a three-layer sentence representation. In DocEng, 2014.

[13] R. Mihalcea, C. Corley, and C. Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In AAAI, 2006.