Short Text Similarity with Word Embeddings (abstract)

Tom Kenter
tom.kenter@uva.nl

Maarten de Rijke
derijke@uva.nl

University of Amsterdam, Amsterdam, The Netherlands

ABSTRACT

Determining semantic similarity between two texts is to find out if two pieces of text mean the same thing. Being able to do so successfully is beneficial in many settings in information retrieval like search [7], query suggestion [8], automatic summarization [1] and image finding [3].

In the present work we aim for a generic model that requires no prior knowledge of natural language, such as parse trees, and no external resources of structured semantic information, like Wikipedia or WordNet.

Recent developments in distributional semantics, in particular neural network-based approaches like [9, 11] only require a large amount of unlabelled text data. This data is used to create a, so-called, semantic space. Terms are represented in this semantic space as vectors that are called *word embeddings*. The geometric properties of this space prove to be semantically and syntactically meaningful [4, 9–11], that is, words that are semantically or syntactically similar tend to be close in the semantic space.

A challenge for applying word embeddings to the task of determining semantic similarity of short texts is going from word-level semantics to short-text-level semantics. This problem has been studied extensively over the past few years [2, 6, 12].

In this work we propose to go from word-level to short-text-level semantics by combining insights from methods based on external sources of semantic knowledge with word embeddings. In particular, we perform semantic matching between words in two short texts and use the matched terms to create a saliency-weighted semantic network. A novel feature of our approach is that an arbitrary number of word embedding sets can be incorporated, regardless of the corpus used for training, the underlying algorithm, its parameter settings or the dimensionality of the word vectors. We derive multiple types of meta-features from the comparison of the word vectors for short text pairs and from the vector means of their respective word embeddings, that have not been used before for the task of short text similarity matching.

We show on a publicly available test collection that our generic method, that does not rely on external sources of structural semantic knowledge, outperforms baseline methods that work under the same conditions and outperforms all methods, to our knowledge, that do use external knowledge bases and that have been evaluated on this dataset.

A full version of this paper was presented at CIKM'15 [5].

1. REFERENCES

[1] R. M. Aliguliyev. A new sentence similarity measure and sentence based extractive technique for automatic text summarization. *Expert Systems with Applications*, 2009.

[2] P. Annesi, D. Croce, and R. Basili. Semantic compositionality in tree kernels. In *CIKM 2014*, 2014.

[3] T. A. Coelho, P. P. Calado, L. V. Souza, B. Ribeiro-Neto, and R. Muntz. Image retrieval using multiple evidence ranking. *IEEE Transactions on Knowledge and Data Engineering*, 16 (4):408–417, 2004.

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

[5] T. Kenter and M. de Rijke. Short text similarity with word embeddings. In *Proceedings of the 24th ACM international conference on information and knowledge management. In CIKM*, volume 15, page 115, 2015.

[6] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

[7] H. Li and J. Xu. Semantic matching in search. *Foundations and Trends in Information Retrieval*, 7(5):343–469, 2014.

[8] D. Metzler, S. Dumais, and C. Meek. Similarity measures for short segments of text. In *ECIR 2007*, 2007.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS 2013*, 2013.

[11] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. *EMNLP 2014*, 2014.

[12] R. Socher, E. H. Huang, J. Pennin, C. D. Manning, and A. Y. Ng. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *NIPS 2011*, 2011.