Siamese CBOW: Optimizing Word Embeddings for Sentence Representations [ABSTRACT]

Tom Kenter[1] tom.kenter@uva.nl
Alexey Borisov[1, 2] alborisov@yandex-team.ru
Maarten de Rijke[1] derijke@uva.nl

[1] University of Amsterdam, Amsterdam
[2] Yandex, Moscow

We present Siamese CBOW, *Siamese Continuous Bag of Words*, a neural network for efficient estimation of high-quality sentence embeddings.[1] Averaging the embeddings of words in a sentence has proven to be a surprisingly successful and efficient way of obtaining sentence embeddings. However, word embeddings trained with the methods currently available are not optimized for the task of sentence representation, and, hence, likely to be suboptimal. Siamese CBOW handles this problem by training word embeddings directly for the purpose of being averaged. The underlying neural network learns word embeddings by predicting, from a sentence representation, its surrounding sentences. We show the robustness of the Siamese CBOW model by evaluating it on 20 datasets stemming from a wide variety of sources.

Word embeddings have proven to be beneficial in a variety of tasks in NLP such as machine translation [13], parsing [1], semantic search [10, 12], and tracking the meaning of words and concepts over time [4, 6]. It is not evident, however, how word embeddings should be combined to represent larger pieces of text, like sentences, paragraphs or documents. Surprisingly, simply averaging word embeddings of all words in a text has proven to be a strong baseline or feature across a multitude of tasks [2, 3].

Word embeddings, however, are not optimized specifically for representing sentences. In this paper we present a model for obtaining word embeddings that are tailored specifically for the task of averaging them. We do this by directly including a comparison of sentence embeddings—the averaged embeddings of the words they contain—in the cost function of our network.

Word embeddings are typically trained in a fast and scalable way from unlabeled training data. As the training data is unlabeled, word embeddings are usually not task-specific. Rather, word embeddings trained on a large training corpus, like the ones from [9] are employed across different tasks [3, 11]. These two qualities — (1) being trainable from large quantities of unlabeled data in a reasonable amount of time, and (2) robust performance across different tasks— are highly desirable and allow word embeddings to be used in many large-scale applications. In this work we aim to optimize word embeddings for sentence representations in the same manner. We want to produce general purpose sentence embeddings that should score robustly across multiple test sets, and we want to leverage large amounts of unlabeled training material.

In the word2vec algorithm, [8] construe a supervised training criterion for obtaining word embeddings from unsupervised data, by predicting, for every word, its surrounding words. We apply this strategy at the sentence level, where we aim to predict a sentence from its adjacent sentences [7]. This allows us to use unlabeled training data, which is easy to obtain; the only restriction is that documents need to be split into sentences and that the order between sentences is preserved.

The main research question we address is whether directly optimizing word embeddings for the task of being averaged to produce sentence embeddings leads to word embeddings that are better suited for this task than word2vec does. Therefore, we test the embeddings in an unsupervised learning scenario. We use 20 evaluation sets that stem from a wide variety of sources (newswire, video descriptions, dictionary descriptions, microblog posts). Furthermore, we analyze the time complexity of our method and compare it to our baselines methods.

Summarizing, our main contributions are:

- We present Siamese CBOW, an efficient neural network architecture for obtaining high-quality word embeddings, directly optimized for sentence representations;
- We evaluate the embeddings produced by Siamese CBOW on 20 datasets, originating from a range of sources (newswire, tweets, video descriptions), and demonstrate the robustness of embeddings across different settings.

1. REFERENCES

[1] D. Chen and C. D. Manning. A fast and accurate dependency parser using neural networks. In *EMNLP*, 2014.

[2] S. J. Gershman and J. B. Tenenbaum. Phrase similarity in humans and machines. In *CogSci*, 2015.

[3] T. Kenter and M. de Rijke. Short text similarity with word embeddings. In *CIKM*, 2015.

[4] T. Kenter, M. Wevers, P. Huijnen, and M. de Rijke. Ad hoc monitoring of vocabulary shifts over time. In *CIKM*, 2015.

[5] T. Kenter, A. Borisov, and M. de Rijke. Siamese cbow: Optimizing word embeddings for sentence representations. In *ACL*, 2016.

[6] Y. Kim, I. Yi-Chiu., K. Hanaki, D. Hegde, and S. Petrov. Temporal analysis of language through neural language models. *ACL*, 2014.

[7] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. Skip-thought vectors. In *NIPS*, 2015.

[8] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *ICLR*, 2013.

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

[10] R. Reinanda, E. Meij, and M. de Rijke. Mining, ranking and recommending entity aspects. In *SIGIR*, 2015.

[11] R. Socher, B. Huval, C. D. Manning, and A. Y. Ng. Semantic compositionality through recursive matrix-vector spaces. In *EMNLP-CoNLL*, 2012.

[12] N. Voskarides, E. Meij, M. Tsagkias, M. de Rijke, and W. Weerkamp. Learning to explain entity relationships in knowledge graphs. In *ACL-IJCNLP*, 2015.

[13] W. Y. Zou, R. Socher, D. M. Cer, and C. D. Manning. Bilingual word embeddings for phrase-based machine translation. In *EMNLP*, 2013.

[1] This abstract is based on [5].