Corpus Size	Word Analogy				Word Similarity (average results on six testing datasets)			
	CBOW-p	CBOW-a	WordRank	OptRank	CBOW-p	CBOW-a	WordRank	OptRank
128M	0.364	0.404	0.415	0.437	0.622	0.618	0.633	0.637
256M	0.438	0.513	0.518	0.542	0.634	0.621	0.651	0.654
512M	0.543	0.632	0.642	0.658	0.643	0.637	0.657	0.675
1G	0.660	0.667	0.647	0.675	0.641	0.631	0.670	0.661
2G	0.691	0.712	0.685	0.718	0.647	0.646	0.665	0.672

Table 2: The best performance of each word embedding model in two testing tasks when the training datasets are relatively small

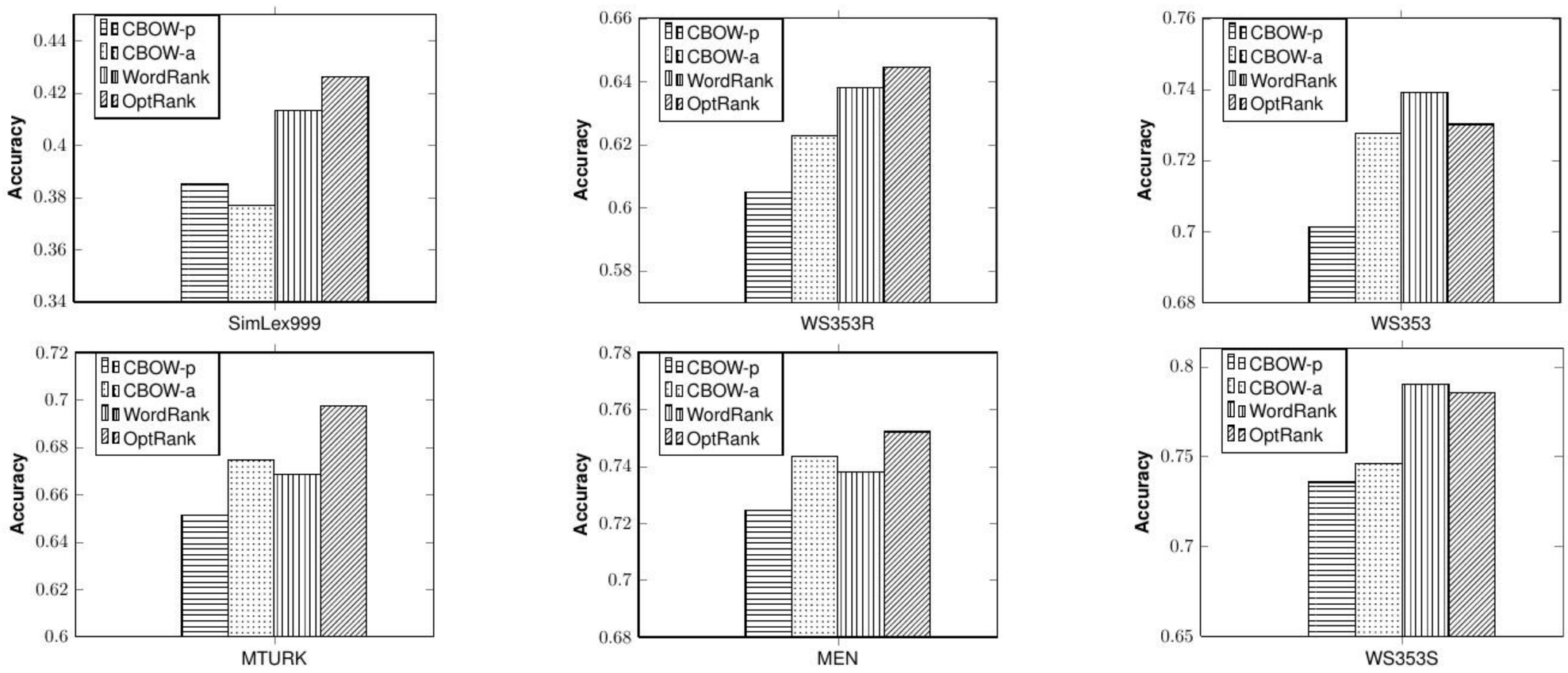

Figure 2: The best performance of each word embedding model (trained on 14G Wiki2017) for the task of word similarity

Model	Semantic	Syntactic	Overall
CBOW-p	0.782	0.678	0.725
CBOW-a	0.811	0.694	0.747
WordRank	0.775	0.687	0.722
OptRank	0.824	0.698	0.756

Table 3: The best perfomance of comparison models (trained on 14G Wiki2017) for the task of word analogy with neg = 2

For the task of word similarity, Figure 2 shows that the OptRank model consistently yields a much better performance than the CBOW-p and CBOW-a model, and beats WordRank on many datasets. Table 3 shows that OptRank is dominant on the word analogy task for all cases. We can observe that CBOW-a performs better than CBOW-p which is in turn better than WordRank. The reason is that WordRank only focuses on the ranks of positive words, but pays no attention to their differences relative to negative words.

5 Conclusion and Future Work

In this paper, we view word embedding as a ranking problem and then analyze the main disadvantage of CBOW model that it does not consider the relation between positive and negative words. This easily results in incorrect ranks of words, and produces suboptimal embeddings during training. Thus, we proposed a novel rank model which learns word representations not only by weighting positive words, but also by oversampling informative negative words. Other models typically only pay attention to one of them. Moreover, by using an effectively learning scheme, we reduce the computational cost of the OptRank, which makes it become a more practising model. These attributes significantly enable OptRank to achieve good performance even if the training datasets are limited.

Although our idea can be directly applied to the skip-gram model, the empirical study shows that the improvement is not as stable as CBOW. The reason is that there is only one target (positive) word in CBOW, but a set of positive words in skip-gram. Hence, in the future we intend to investigate how to handle the scenario with a set of target words. Meanwhile, we are also interested to compare our OptRank with a newly proposed embedding model Allvec [Xin *et al.*, 2018], which is learned by batch gradient descent with all negative examples instead of SGD with negative sampling.

Acknowledgments

This work was supported by the National Natural Science Foundation for Young Scientists of China under Grant No. (61702084, 61772125, 61702090) and the Fundamental Research Funds for the Central Universities under Grant No.N161704001.