Approximating Word Ranking and Negative Sampling for Word Embedding

Guibing Guo, Shichang Ouyang, Fajie Yuan, Xingwei Wang
{guogb,wangxw}@swc.neu.edu.cn, 1701282@stu.neu.edu.cn, f.yuan.1@research.gla.ac.uk

Introduction

- CBOW (Continuous Bag-Of-Words) is one of the most commonly used techniques to generate word embeddings in various NLP tasks. However, it fails to reach the optimal performance due to uniform involvements of positive words and a simple sampling distribution of negative words.
- We formalize word embedding as a ranking problem and propose a new model called OptRank which weighs the positive words by their ranks such that highly ranked words have more importance, and adopts a dynamic sampling strategy to select informative negative words.
- In addition, an approximation method is designed to efficiently compute word ranks. Empirical experiments show that OptRank consistently outperforms its counterparts on a benchmark dataset with different sampling scales, especially when the sampled subset is small.

OptRank Model

Weigh Positive Word: We weigh the positive word w_p according to its rank which is the same as the number of words that have greater relevant scores with context c than w_p, given by:

$$rank(w_p, c) = \sum_{w \in W} I(v_c^\top v_{w_p} < v_c^\top v_w + \varepsilon) \tag{1}$$

Our objective is to minimize the rank value of positive words, and formulated as follows.

$$\mathcal{O}_{(w_p,c)} = f(rank(w_p, c)) = \log_2(rank(w_p, c)). \tag{2}$$

Dynamic Sampling Strategy: we intend to select the 'informative' negative words that are likely to mess up our model. Specifically, we opt to select the negative words that satisfy the following requirement:

$$v_c^\top v_{w_n} + \varepsilon > v_c^\top v_{w_p} \tag{3}$$

Loss Function: Combing positive ranking and negative sampling together, we can obtain the following objective function to maximize the classification probability.

$$\begin{aligned}\mathcal{J} = \ & \textstyle\sum_{(w,c)} \Big\{ \mathcal{O}_{(w_p,c)} \cdot \{ -\log(\sigma(v_c^\top v_{w_p})) \} \\ & + \textstyle\sum_{w_n \in N} \{ -\log(1 - \sigma(v_c^\top v_{w_n})) \} \Big\}\end{aligned} \tag{4}$$

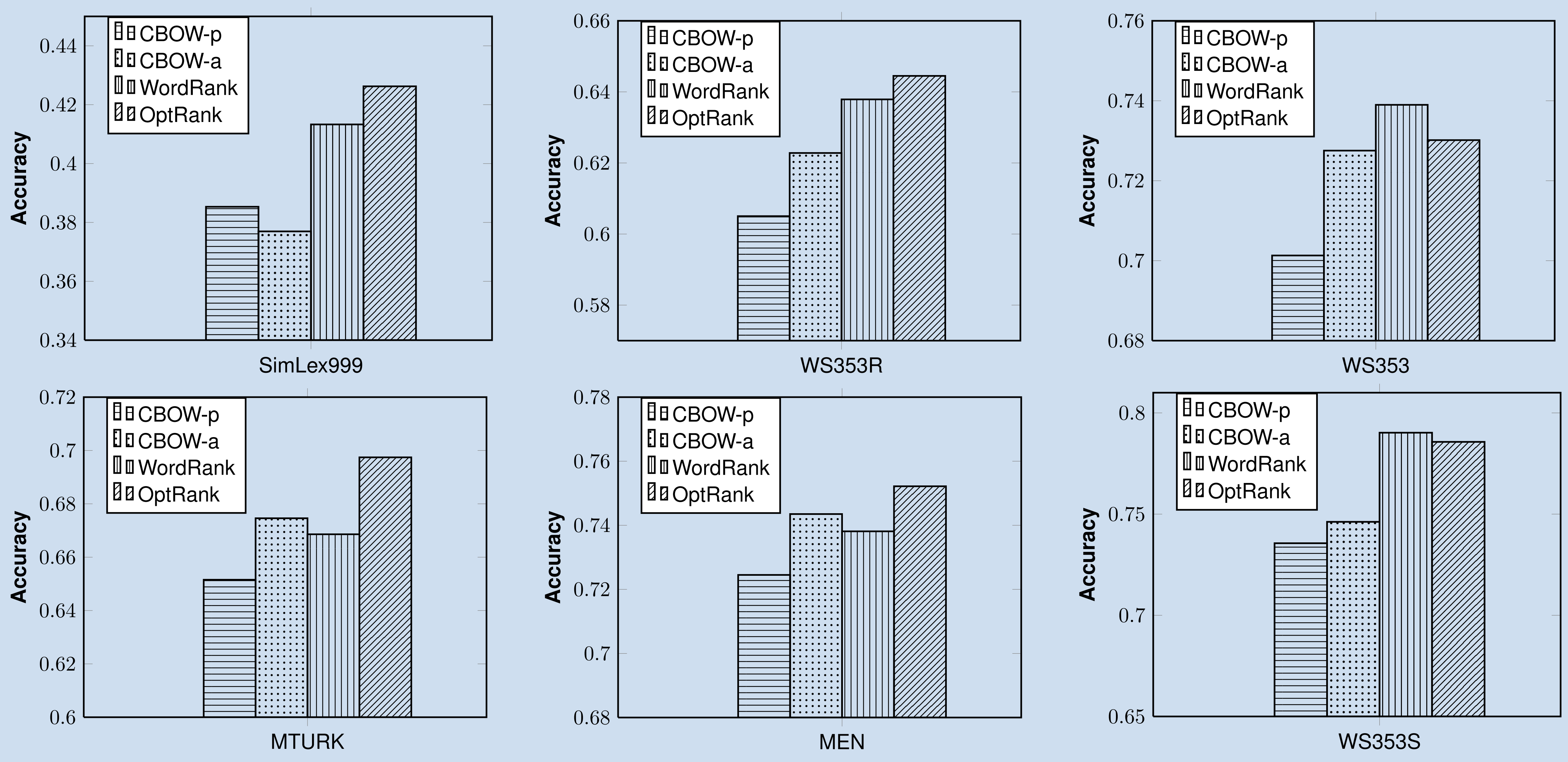
Effective Learning Scheme: Computing the exact value of $rank(w_p, c)$ requires an exhaustive search in the whole word space. It is a very time-consuming step. Thus, we repeatedly sample a negative word from the corpus W until we obtain an expected word w_n that satisfy the requirement given by Eq. 3 to estimate the $rank(w_p, c)$.

Experimental Results

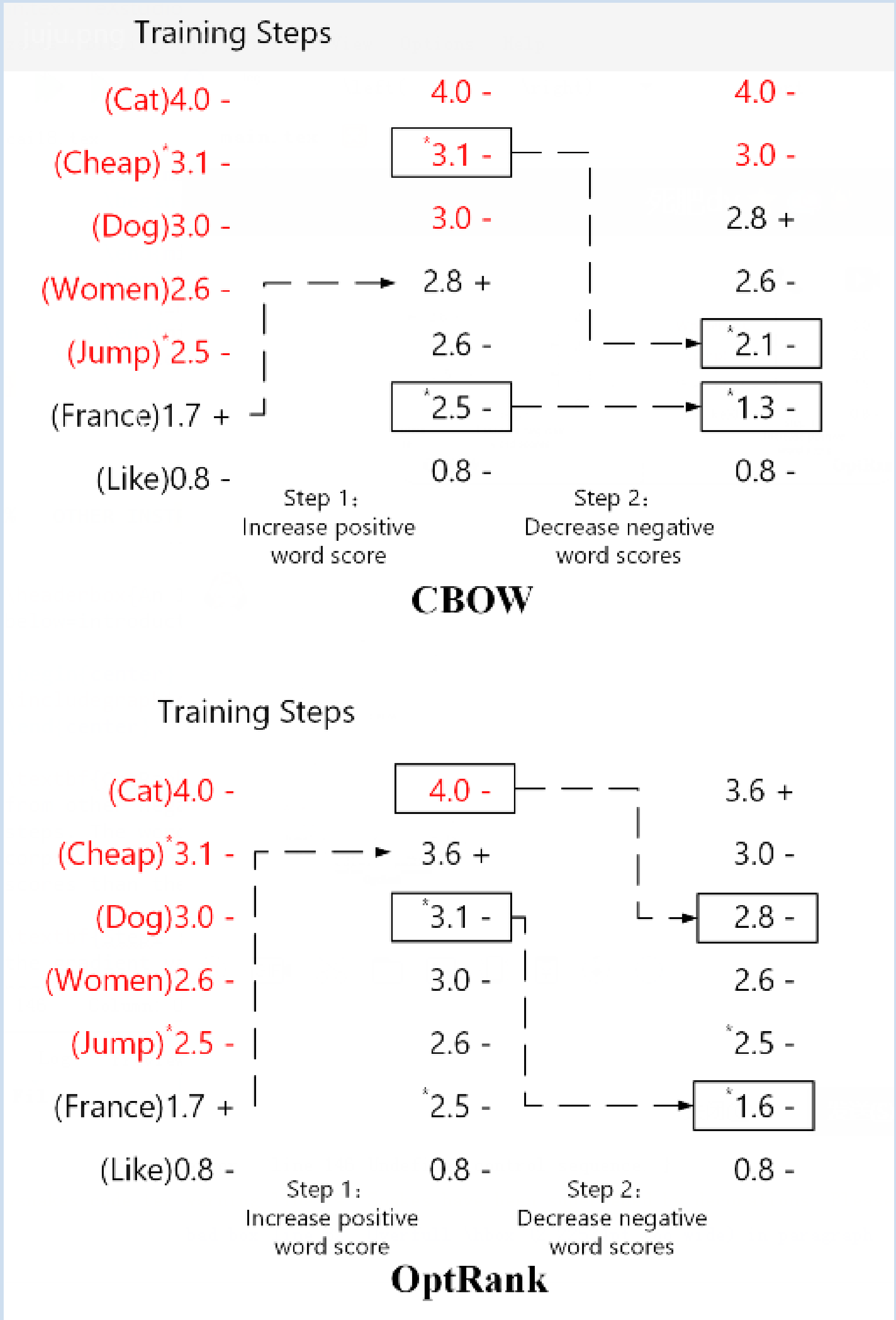
The word analogy task is to answer the questions in the form of "a is to b as c is to ?" and the word similarity task is to calculate the consine similarity between two relevant words.

Corpus	Word Analogy			Word Similarity		
	CBOW	WordRank	OptRank	CBOW	WordRank	OptRank
128M	0.364	0.415	0.437	0.622	0.633	0.637
256M	0.438	0.518	0.542	0.634	0.651	0.654
512M	0.543	0.642	0.658	0.643	0.657	0.675
1G	0.660	0.647	0.675	0.641	0.670	0.661
2G	0.691	0.685	0.718	0.647	0.665	0.672

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



An Intuitive Example



OptRank vs CBOW In order to distinguish the positive word 'France'(+) from other negative words(−). Both models will train the word vector in two steps. The word denoted with symbol '*' means that it is a popular word in the corpus. And the red words are the negative words which have higher relevance scores than the positive word.

Step1 : CBOW will increase the relevance score of the positive word by the gradient values from 1.7 to 2.8, but the OptRank model can increase the ranking score of the positive word to a larger extent with the help of item ranks.

Step2 : CBOW will sample some popular words (e.g., Cheap) denoted by '*' as the negative words, leading to a better yet suboptimal ranking list, but OptRank adopts dynamic sampling to find an informative negative example (i.e., word cat) and decrease its ranking score.

Conclusion

We view word embedding as a ranking problem and then analyze the main disadvantage of CBOW model. Then, we proposed a novel rank model which learns word representations not only by weighting positive words, but also by oversampling informative negative words.

Contact

English Page

Wechat