Distribution-Based Chinese Word Representation Research (Postprint)

Authors: Cao Xuefei, Li Jihong, Wang Ruibo

Date: 2018-05-18T00:00:00+00:00

Abstract

We conducted a systematic investigation into the parameter selection problem in constructing distribution-based Chinese word representations. Six parameters were selected for comparative experiments, and the quality of the resulting Chinese word representations under different parameter settings was assessed on Chinese semantic similarity tasks. The experimental results indicate that with appropriate parameter choices, distribution-based word representations can achieve high performance on Chinese semantic similarity tasks. Furthermore, the quality of these high-dimensional distributional word representations even outperforms that of the currently popular low-dimensional word representations obtained through neural network-based (Skip-gram) or matrix factorization-based (GloVe) methods.

Full Text

Preamble

Study of Distributional Representation of Chinese Words

Cao Xuefei, Li Jihong, Wang Ruibo
(School of Software, Shanxi University, Taiyuan 030006, China)

Abstract: This paper presents a systematic investigation into parameter selection for constructing distributional representations of Chinese words. We selected six types of parameters for comparative experiments and evaluated the quality of Chinese word representations obtained under different parameter settings on a Chinese semantic similarity task. Experimental results demonstrate that by choosing appropriate parameters, distributional word representations can achieve high performance on Chinese semantic similarity tasks. Moreover, the quality of such high-dimensional distributional representations even surpasses that of low-dimensional word representations obtained from currently popular neural network (Skip-gram) or matrix factorization (GloVe) methods.

Keywords: distributional representation; semantic similarity; pointwise mutual information

0 Introduction

Words serve as the fundamental units of natural language and the basic carriers of semantic information. Formalizing words into symbolic representations that encapsulate their semantic content is foundational to natural language understanding and processing. Current word representation methods primarily fall into two categories: count-based methods and prediction-based methods. Count-based methods, also known as distributional representation methods, rely on the distributional hypothesis: words that appear in similar contexts share similar semantics. Consequently, modeling the association between words and their contexts in a corpus can capture word meanings. Co-occurrence frequency is commonly employed as the association measure. Specifically, the Vector Space Model (VSM) maps words to vectors in a semantic space, where each dimension corresponds to a context word and each element value represents the co-occurrence frequency within a certain window. All word vectors in the lexicon can be organized into a matrix known as the word-context co-occurrence matrix.

For specific tasks, constructing an appropriate VSM heavily depends on parameters such as context definition, window type and size, with different parameter settings significantly impacting model performance. For instance, reference [3] examined the relationship between model performance and vector dimensionality, showing that on semantic similarity tasks, word representation vectors with 50K dimensions yielded better results than those with 1K dimensions. Reference [6] systematically studied various VSM parameters—including context window type and size, vector dimensionality, corpus size, and weighting methods for co-occurrence frequency—on four semantic tasks, concluding that using PMI (pointwise mutual information) for weighting, the smallest possible context windows, and the largest possible vector dimensions produced the best word representations. Reference [7] investigated seven types of parameters across different English corpora and conducted a series of comparative experiments on semantic similarity tasks, obtaining several empirical findings: (a) results stabilized as vector dimensionality increased; (b) among different weighting methods, PMI was the superior choice; and (c) corpus type and size substantially affected results, with larger corpora of the same type yielding better performance.

In summary, no consensus currently exists regarding parameter selection and configuration in VSM, and the aforementioned work has focused exclusively on English. To our knowledge, no systematic study has addressed parameter selection for distributional Chinese word representations. Therefore, this research focuses on two questions: (a) Can distributional representation methods be applied to Chinese, and how do they perform (using Chinese semantic similarity tasks as examples)? (b) Do parameters used in English distributional semantic models exhibit consistent behavior in Chinese? Based on these questions, we conducted detailed comparative experiments to explore parameter selection for

Chinese word representations.

1 Parameters for Constructing Chinese Word Distributional Representations

VSM can represent words as vectors in semantic space, with the advantage of enabling linear algebraic operations to compute distances between vectors and thereby determine word similarity. However, VSM construction involves numerous parameters, and parameter selection often determines final model performance.

1.1 Weighting Methods for Word-Context Co-Occurrence Frequency

Co-occurrence frequency refers to the number of times a word and its context (words appearing around the target word) co-occur within a certain window size. Representing a word as a vector where each dimension corresponds to a context and each element value represents the co-occurrence frequency yields a distributional word representation. However, research indicates that directly using raw co-occurrence frequencies as vector representations is ineffective. Quality can be improved through mathematical transformations, such as reweighting the co-occurrence matrix using different methods. Common weighting approaches include PMI, t-score, and log-likelihood ratio, with PMI being one of the most frequently used methods. PMI is defined as:

$$P(w,c)\log\frac{P(w,c)}{P(w)P(c)}$$

where $P(w,c)$ represents the probability of word w and context c co-occurring, and $P(w)$ and $P(c)$ represent the probabilities of word w and context c appearing in the corpus, respectively. Note that when $P(w,c)=0$, the PMI value is negative infinity; to avoid this, the PMI value is defined as 0 in such cases.

Additionally, we consider a PMI variant, Positive PMI (PPMI), as a weighting method for co-occurrence frequency. PPMI is defined as:

$$\mathrm{PPMI}(w,c)=\max(0,\mathrm{PMI}(w,c))$$

Reference [6] demonstrates that using PPMI for word co-occurrence frequency weighting outperforms PMI on English semantic similarity tasks.

1.2 Context Window Type and Size

Co-occurrence frequency is typically 统计 based on context windows, taking a certain number of words to the left and right of the target word as context. If each word has a fixed constant window size, we call this the fixed window method. Reference [10] proposed a dynamic context window approach: setting

a window threshold parameter (thr), generating a random number r from the interval $[1, thr]$ for each word when constructing its context, and then taking r words before and after the target word as context. Under this method, the number of context words for each target word is randomly generated, which we term the random window method.

Research shows that context window size significantly impacts the performance of distributional word representations across different tasks. Reference [11] notes that smaller windows should be used for syntax-related tasks, while larger windows are recommended for semantics-related tasks. Reference [12] found that smaller windows should be used to measure similarity between concrete nouns, while larger windows are preferable for abstract nouns. To investigate appropriate window settings for Chinese semantic similarity tasks, we set windows to 1, 2, 5, 8, 10, and 12 in our experiments. Under the fixed window method, we take the specified number of words to the left and right as context, while under the random window method, these values serve as window threshold parameters.

1.3 Context Smoothing Coefficient

When calculating PMI values, frequency information 统计 from the corpus can be used to estimate corresponding probability values. For example, the probability of a context can be estimated using:

$$P(c) = \frac{\#(c)}{\sum_{c'} \#(c')}$$

where $\#(c)$ represents the count of context c in the corpus. However, directly using this formula presents a problem: PMI values tend to be "biased" toward rare words. For certain words with very low frequency, the PMI value calculated based on the formula can become extremely large. Consequently, existing research introduces a smoothing coefficient α to smooth the context distribution:

$$P(c) = \frac{\#(c)^\alpha}{\sum_{c'} \#(c')^\alpha}$$

This smoothing also leads to changes in PMI calculation:

$$\mathrm{PMI}_\alpha(w, c) = \log \frac{P(w, c)}{P(w)P(c)^\alpha}$$

In our experiments, we set the smoothing coefficient to 1 and 0.75 (based on existing English experiments).

1.4 PMI Shift Parameter (neg)

It is generally believed that negative PMI values may not help improve the quality of distributional word representations. Therefore, some English-based distributional word representation methods directly use PPMI instead of PMI. Moreover, using PPMI makes the co-occurrence matrix sparse, which facilitates computation. Reference [9] further generalized PPMI:

$$\text{SPPMI}(w, c) = \max(0, \log \frac{P(w, c)}{P(w)P(c)} - \log(\text{neg}))$$

where neg is called the shift parameter. When neg = 1, SPPMI is equivalent to PPMI. When neg > 1, it retains the mutual information of word-context pairs with high association in the co-occurrence matrix while increasing matrix sparsity. When neg < 1, the co-occurrence matrix includes mutual information of some unrelated co-occurring word pairs and reduces matrix sparsity. Similarly, the shift parameter neg can be applied to PMI:

$$\text{SPMI}(w, c) = \text{PMI}(w, c) - \log(\text{neg})$$

Thus, we obtain two additional PMI variants: SPMI and SPPMI. In experiments, we empirically set neg values to 0.2, 0.5, 0.8, 2, 5, and 8.

1.5 High-Frequency Word Subsampling (sub)

In a large corpus, some high-frequency words appear hundreds of thousands or even millions of times, yet provide little useful information (e.g., “是” [is], “的” [possessive particle]). One could simply remove these high-frequency words from the corpus, but this would cause the context windows of many words adjacent to them to expand, making some word pairs co-occur that would not have in the original corpus. Reference [10] proposed a subsampling technique that accelerates computation while improving word representation quality to some extent. The subsampling method works as follows: set a threshold t, and when scanning the corpus, high-frequency words with frequency greater than t will be ignored with probability:

$$P(f(w)) = 1 - \sqrt{\frac{t}{f(w)}}$$

where $f(w)$ is the frequency of word w in the corpus. In our experiments, based on the actual corpus used, we set t to 10^{-3}, 0.5×10^{-3}, and 10^{-4}.

1.6 Vector Dimensionality

Each vector element in a word's distributional representation corresponds to one context of that word. Vector dimensionality significantly impacts distributional word representations. How should vector dimensionality be set for Chinese word distributional representations? That is, how many contexts should be selected to represent a word, or can certain types of words be chosen as contexts? In our experiments, we sorted words by frequency in descending order to serve as representation dimensions and selected different dimension sizes and word types to examine how heavily the quality of distributional representations depends on vector dimensionality.

2 Corpus and Test Datasets

We selected the People' s Daily corpus annotated by Peking University (including 1998 and 2000) as our experimental corpus. After removing annotation information, the corpus contains approximately 29 million words.

Accurately interpreting a word' s semantic information requires background knowledge or contextual information, making it difficult to directly evaluate the quality of a word representation. There is no widely accepted evaluation metric or standard dataset for directly assessing distributional representation quality; only indirect evaluation from certain perspectives is possible. We selected Chinese semantic similarity as our evaluation task, which determines semantic similarity between two words by calculating the cosine similarity between their representation vectors. We used wordsim297 as our test dataset, which contains 297 word pairs, each with human-annotated similarity scores (higher scores indicate greater semantic similarity). We used vector cosine similarity as the metric for semantic similarity in the dataset and Spearman' s rank correlation coefficient to measure the correlation between human annotations and computed results.

3 Experimental Results

Due to the large number of parameters involved in the experiments, we adopted a simple strategy: first analyzing some parameters, then fixing those values to study other parameters. In our experiments, we first determined window type and size, then examined how other parameters affected distributional word representations.

3.1 Impact of Context Window

We first examined the performance of distributional word representations obtained from PMI and its three variants (PPMI, SPMI, and SPPMI) under different context window types and sizes on the Chinese semantic similarity task. SPMI and SPPMI were set with neg = 2, and contexts consisted of all words appearing in the corpus (vector dimensionality approximately 300K). Figure

1 shows the experimental results. Under the fixed window method, all four weighting methods essentially achieved optimal distributional word representations with a window size of 5, and results stabilized when the window exceeded 5. Under the random window method, PMI, PPMI, and SPPMI achieved optimal window values of 8, with experimental results slightly decreasing when the window exceeded 8. For SPMI, although results improved slightly when the window was set to 12, the improvement was not significant. In Figure 1, the X-axis represents window size, and the Y-axis represents Spearman' s rank correlation coefficient ($\times$1000).

Figure 2 compares the performance of fixed and random window methods across different window sizes and weighting methods. Results show that under any weighting method, the fixed window method consistently outperforms the random window method. Even when the random window method achieves its best performance at window size 8, it remains inferior to the fixed window method at window size 5. Therefore, in subsequent experiments, we set the context window type to fixed with a size of 5.

3.2 Impact of Context Smoothing Coefficient

Figure 3 presents experimental results for context smoothing coefficients across different weighting methods and vector dimensionalities. The X-axis represents vector dimensionality (in thousands), and the Y-axis represents Spearman' s rank correlation coefficient ($\times$1000). We can clearly draw two conclusions:

(a) Smoothing the context distribution (cds = 0.75) does not improve the performance of distributional word representations on Chinese semantic similarity tasks.

(b) For Chinese word distributional representations, there is no need to set larger vector dimensionalities. When dimensionality exceeds 130K, performance does not significantly improve (Figure 3(b) even shows a slight decrease). However, when dimensionality is less than 130K and gradually increases, performance improves noticeably.

In our experiments, a dimensionality of 130K corresponds to contexts that appeared at least 10 times in the corpus. The figure shows that removing low-frequency words (words appearing fewer than 10 times) from contexts does not affect distributional word representation performance. While some English corpus experiments suggest that larger dimensionalities yield better performance on semantic similarity tasks, our Chinese results indicate that not all words need to be used as contexts. When dimensionality reaches a certain number, stable results can be obtained on semantic similarity tasks; further increasing dimensionality produces no significant changes but increases computational overhead.

3.3 Impact of Shift Parameter

Applying the shift parameter neg to co-occurrence frequencies after PMI and PPMI weighting, we found that an appropriate shift parameter neg can improve the performance of PPMI-based distributional representations, but this parameter has no effect on PMI. Figure 4(a) shows that as the neg value gradually increases, the performance of distributional word representations on semantic similarity tasks also improves, reaching maximum performance when neg = 2. Figure 4(b) shows that whether neg is greater than 1 or less than 1, the results are smaller than when neg = 1. We believe that compared to PMI, PPMI removes co-occurrence information for word-context pairs with negative mutual information in the co-occurrence matrix, retaining only context information with certain associations to the word. That is, for a word' s distributional representation, we filter out contexts not associated with the word, and an appropriate shift parameter further filters out contexts with low association (as in Figure 4(a) when neg = 2). However, if the shift parameter is too large, it may discard context information strongly associated with the word, thereby reducing distributional representation quality (as shown in Figure 4(b) when neg takes larger values of 5 and 8, results 反而 decrease).

3.4 Impact of High-Frequency Word Subsampling

Subsampling technology can dilute high-frequency words in the corpus, similar to removing stop words. Experimental results show that subsampling does not improve the performance of distributional representations generated by PPMI and SPPMI weighting methods, but provides clear benefits for PMI and SPMI. As shown in Figures 5(c-d), when the sampling parameter sub is set to 10^{-4}, the distributional representations of PMI and SPMI 统计 using subsampling technology both show improved performance on Chinese semantic similarity tasks. We believe this phenomenon occurs because applying subsampling technology to PMI and SPMI weighting methods effectively increases the window size to some extent, raising the probability of certain word pairs co-occurring in the corpus (i.e., increasing the numerator in Equation (1)), causing some word pairs' PMI (SPMI) values to change from negative to positive. This causes the PMI matrix to approach a PPMI matrix, and as previously mentioned, PPMI outperforms PMI. However, it should be noted that even with subsampling, the performance improvement of PMI and SPMI is insufficient to close the gap with SPPMI (Figure 5). Overall, among the four weighting methods, SPPMI is optimal (see Figures 1-3).

3.5 Vector Dimensionality Selection

Previous experimental results show that we do not need all words as contexts (see Section 3.2). This is partly because word frequencies in corpora follow Zipf's law, and removing some low-frequency words does not significantly affect results. As our experiments demonstrate, selecting words that appeared more than 10 times as contexts (approximately 130K words) yields good and stable distribu-

tional word representations. In addition to removing low-frequency words, we also experimented with other methods for setting vector dimensionality, such as removing stop words or selecting words within certain frequency ranges, but these did not produce significant performance improvements.

Considering that Chinese words can be divided into content words and function words based on grammatical function, we selected nouns, verbs, and adjectives from content words as contexts, using PMI as the co-occurrence frequency weighting method, a context window of 5, without subsampling or context smoothing. Table 1 shows the experimental results.

Table 1 Experimental results using content words as contexts

Context Type	Context Dimensionality	Result
Words appearing >10 times in corpus	~130K dimensions	
Frequent nouns, verbs, and adjectives among content words	~5,500 dimensions	

Using fixed context window, size = 5

The results show that using content words as contexts not only reduces vector dimensionality, thereby simplifying subsequent vector operations, but also improves word representation performance on Chinese semantic similarity tasks to some extent. This is because, compared to function words, content words carry substantive meaning, and each content word can be explained in detail. Using content words as contexts better represents word meanings.

3.6 Comparison with Other Methods

Based on the above experimental results, we can derive the following empirical findings: When constructing Chinese word distributional representations using the vector space model, we recommend using a fixed-size window (window size can be 尝试 set to 5); applying PPMI weighting to co-occurrence frequencies and setting a shift parameter greater than 1 (e.g., 2 in our experiments); and we do not recommend other parameters or techniques used in English, such as context smoothing coefficients and high-frequency word subsampling technology.

Additionally, we compared our high-dimensional distributional representation (SPPMI) obtained under the best parameter configuration with low-dimensional dense representations learned from currently popular neural network-based word2vec (CBOW and Skip-gram) and matrix factorization-based GloVe methods on Chinese semantic similarity tasks. Table 2 shows CBOW, Skip-gram, and GloVe results from reference [15], trained on the same corpus as ours with vector dimensionality of 200. Experimental results indicate that on

Chinese similarity tasks, PPMI-based distributional word representations are comparable to these methods. Although the results are lower than CBOW, they show clear improvement over Skip-gram and GloVe, confirming the conclusion in reference [9] that with appropriate parameter settings, the quality of high-dimensional distributional word representations is not inferior to low-dimensional representations (embeddings) learned through neural network or other methods.

Table 2 Comparison of our best results (SPPMI) with other methods

Using fixed context window, size = 5

4 Conclusion

This paper conducted a systematic study of parameters involved in Chinese word distributional representations and provided practical recommendations for parameter selection and configuration based on experimental results. Specifically, for window type selection, fixed-size windows outperform dynamic windows, and window size should not be too small—we recommend setting it to 5 (i.e., 5 words on each side). Raw co-occurrence frequencies require weighting (first PPMI weighting, then reasonable shift parameter setting). For vector dimensionality (i.e., number of contexts), one can simply use all words as contexts if computational cost is not a concern, but a better choice is to use only content words as contexts. We do not recommend using context smoothing and subsampling techniques for Chinese word distributional representations.

Our experiments were based on the People' s Daily corpus. Although this manually annotated corpus has high segmentation accuracy, its scale is smaller compared to commonly used English corpora. Different types of larger corpora should be used to validate our results. Additionally, whether dimensionality reduction (e.g., SVD) on high-dimensional Chinese word distributional representations obtained from co-occurrence frequencies can improve representation quality remains to be studied. Finally, our experimental results show that subsampling high-frequency words did not produce the same effects as in English. Whether other preprocessing methods for high-frequency words or certain types of high-frequency words could yield better distributional representations requires further investigation.

References

[1] Baroni M, Dinu G, Kruszewski G. Don' t count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *Proc of Meeting of the Association for Computational Linguistics*, 2014: 238-247.

[2] Harris Z. Distributional Structure. *Word*, 1954: 146-162.

[3] Milajevs D, Sadrzadeh M, Purver M. Robust Co-occurrence Quantification for Lexical Distributional Semantics. *Proc of ACL Student Research Workshop*,

2016: 58-64.

[4] Hanks P, Hanks P. Word association norms, mutual information, and lexicography. *Proc of Meeting on Association for Computational Linguistics*, 1989: 76-83.

[5] Turney, Peter D, Pantel, et al. From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research*, 2010, 37(1): 141-188.

[6] Bullinaria J A, Levy J P. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 2007, 39(3): 510.

[7] Kiela D, Clark S. A Systematic Study of Semantic Vector Space Model Parameters. *Proc of Workshop on Continuous Vector Space Models & Their Compositionality*, 2014: 21-30.

[8] Evert S. The statistics of word cooccurrences: word pairs and collocations. PhD dissertation, University of Stuttgart, 2004.

[9] Levy O, Goldberg Y, Dagan I. Improving distributional similarity with lessons learned from word embeddings. *Bulletin De La Société Botanique De France*, 2015, 75(3): 552-555.

[10] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space. *Computer Science*, 2013.

[11] Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation. *Proc of Conference on Empirical Methods in Natural Language Processing*, 2014: 1532-1543.

[12] Hill F, Kiela D, Korhonen A. Concreteness and corpora: a theoretical and practical analysis. *Proc of the Workshop on Cognitive Modeling and Computational Linguistics*, 2013: 75–83.

[13] Wang Xiang, Jia Yan, Zhou Bin, et al. Semantic relatedness computation based on Chinese Wikipedia link structure and classification system. *Small Microcomputer Systems*, 2011, 32(11): 2237-2242.

[14] Liu Qun, Li Sujian. Word semantic similarity computation based on HowNet. *Chinese Computational Linguistics*, 2002(7): 59-76.

[15] Chen X X, Xu L, Liu Z Y, et al. Joint learning of character and word embeddings. *Proc of International Joint Conference on Artificial Intelligence*, 2015: 1236-1242.

[16] Lebret R, Collobert R. Rehabilitation of count based models for word vector representations. *Computational Linguistics and Intelligent Text Processing*, 2015: 417-429.

Note: Figure translations are in progress. See original paper for figures.