Revisiting Correlations between Intrinsic and Extrinsic Evaluations of Word Embeddings

Yuanyuan Qiu[1,2], Hongzheng Li[3], Shen Li[1,2], Yingdi Jiang[1,2], Renfen Hu[1,2], and Lijiao Yang[1,2]

[1] Institute of Chinese Information Processing, Beijing Normal University
[2] UltraPower-BNU Joint Laboratory for Artificial Intelligence, Beijing Normal University
[3] School of Computer Science & Technology, Beijing Institute of Technology
{reesechiu, lihongzheng, shen, de, irishere}@mail.bnu.edu.cn
yanglijiao@bnu.edu.cn

Abstract. The evaluation of word embeddings has received a considerable amount of attention in recent years, but there have been some debates about whether intrinsic measures can predict the performance of downstream tasks. To investigate this question, this paper presents the first study on the correlation between results of intrinsic evaluation and extrinsic evaluation with Chinese word embeddings. We use word similarity and word analogy as the intrinsic tasks, Named Entity Recognition and Sentiment Classification as the extrinsic tasks. A variety of Chinese word embeddings trained with different corpora and context features are used in the experiments. From the data analysis, we reach some interesting conclusions: there are strong correlations between intrinsic and extrinsic evaluations, and the performance of different tasks can be affected by training corpora and context features to varying degrees.

Keywords: word embedding · intrinsic evaluation · extrinsic evaluation.

1 Introduction

Word embeddings are proved to be beneficial to various Natural Language Processing (NLP) tasks, such as part-of-speech tagging (POS), chunking, named entity recognition (NER), and syntactic parsing[2, 8, 25].

With the increasing usage of word embedding, the issue of evaluation becomes important. Current evaluation methods have two major categories: intrinsic and extrinsic. Intrinsic evaluations directly test for syntactic or semantic relationships between words through word similarity or analogical reasoning tasks [12, 16, 18]. While in extrinsic evaluation, embeddings are exploited as input features for downstream NLP tasks, and their performance can indirectly reflect the effects of embeddings [1]. However, most intrinsic and extrinsic evaluations are conducted separately and few research studies the correlation between them. Chiu et al.[7] argue that most intrinsic evaluations are poor predictors of downstream tasks performance. In their experiment, they compare the embeddings trained with different window sizes on word similarity task and three extrinsic tasks, while another important intrinsic task word analogy is not considered. Moreover,

some effective features such as training corpora and context features have not yet been explored either.

On the other hand, existing discussion of embedding evaluation is mostly about English word embeddings, and there are rich benchmarks in English for both intrinsic and extrinsic evaluation. Although Chinese NLP has grown rapidly in recent years, few attempts have been made in the evaluation of Chinese word embeddings [19].

Based on the above consideration, this paper studies the correlation between intrinsic evaluation and extrinsic evaluation by using 21 Chinese word embeddings trained with different settings. Specifically, we choose word similarity and word analogy as the intrinsic tasks, Named Entity Recognition and Sentiment Classification as the extrinsic tasks. 7 corpora of different sizes and domains are used for training. In addition to the corpus factors, we examine the effectiveness of two important context features during training, i.e. character features and ngram features.

The experimental results demonstrate that both intrinsic and extrinsic performance can be affected by the size and domain of a training corpus, as well as the context features, but influence degrees vary among different tasks. By analyzing the data, we find that there is a consistency between intrinsic and extrinsic evaluations to some extent. Effective features in intrinsic tasks can also improve the performance of extrinsic tasks, but each task may have a preference of specific features. For example, domain-specific corpora have a distinct advantage for extrinsic tasks, and character features are particularly favorable to intrinsic tasks, e.g. analogical reasoning on morphological relations.Thus, this study can not only offer greater and deeper insight on training and evaluating word embeddings, but also some practical suggestions on selecting the suitable word embeddings for NLP tasks.

The contributions of this paper can be summarized as follows: first of all, we present a comprehensive study on the correlation between intrinsic and extrinsic evaluation of word embeddings. We find that intrinsic evaluation can serve as a good predictor for downstream tasks, and different tasks may favor different features. Secondly, we build domain-specific NER and sentiment classification datasets, which could serve as extrinsic benchmarks for evaluation of Chinese word embeddings, as well as other NLP models.

The remaining parts of this paper are organized as follows: section 2 discusses the related work. Section 3 and Section 4 describe the intrinsic and extrinsic tasks respectively. Section 5 conducts experiments and gives analysis in detail. And we give conclusions in Section 6.

2 Related Work

There have been a lot of discussion of the evaluation of word embeddings in recent years. These works study either intrinsic evaluation approaches such as word similarity [4, 11] and word analogy [21], or extrinsic tasks such as POS tagging and Name Entity Recognition. Schnabel et al.[23] present a comprehensive study of intrinsic and extrinsic evaluation of embeddings. Ghannay et al.[13] conduct a detailed comparison of different kinds of word embeddings on various NLP tasks. Relevant works can be found in [14, 22, 28]. As a typical shared task officially proposed in 2002 [24], NER is

one of the downstream tasks commonly used to evaluate the embeddings in most works related to extrinsic evaluations.

However, there are still some challenges and debates in the field of embedding evaluation, for example, Schnabel et al.[23] argue that extrinsic evaluation only provides one way to specify the goodness of an embedding, and it is not clear how it connects to other measures. On the other hand, there are not so many works studying the correlation between intrinsic and extrinsic evaluations. One representative work is [7]. They state that most intrinsic evaluations are poor predictors of downstream tasks performance. However, the experiments consider only one factor in the training of word embeddings, i.e. the window size, a hyper-parameter. Moreover, the intrinsic evaluation only includes the word similarity task, which is insufficient because the effectiveness of word similarity task in evaluation has been questioned a lot, for example, human judgment of word similarity is subjective and similarity is often confused with relatedness [3, 10].

As for evaluation of Chinese word embedding, related work and datasets are much less than that of English. In Chinese, a word is composed of one or more graphical characters, known as Hanzi, which could encode rich semantic and phonetic information. It has attracted considerable attention to use character relevant features to enhance the word representations [6, 20, 27]. To evaluate the newly proposed methods, Chen et al. [6] build a small analogy dataset covering 230 unique Chinese words by translating part of an English dataset. Chen and Ma [5] create several evaluation sets for Chinese word embeddings on both word similarity and analogical tasks. Li et al. [19] release a big and balanced dataset CA8 for analogy evaluation, as well as over 100 Chinese word embeddings trained with different corpora and settings.

Based on previous works, this paper will go further into the evaluation of Chinese word embeddings, and study the correlations between intrinsic and extrinsic evaluation with representative tasks and various embeddings.

3 Intrinsic Tasks

In this paper, we propose to evaluate word embeddings with two representative intrinsic tasks: word similarity and word analogy.

3.1 Word Similarity

Word similarity is an attractive and popular task for embedding evaluation because it is computationally inexpensive and fast. In this task, the correlation coefficient between the automatic predicted results with the human labeled similarity scores is computed. This paper uses the Chinese word similarity dataset proposed by Wu and Li[26].

3.2 Word Analogy

Word analogy task, also called analogical reasoning, aims at detecting morphological and semantic relations between words. Specifically, it is to retrieve the answer of the question “a is to b as c is to ?” with vector computation. We adopt the CA8 dataset constructed by Li et al. [19], including both morphological questions and semantic questions. The questions are solved by 3COSMUL [17] objective.

4 Extrinsic Tasks

To evaluate the performance of word representations in downstream tasks, we apply them to name entity recognition and sentiment analysis. In this paper, we build a Financial NER dataset for name entity recognition, and a Book Review dataset for sentiment classification. These datasets and evaluation methods will be released at Github.

4.1 Named Entity Recognition

Named Entity Recognition (NER) is considered as a typical sequence labeling problem. In NER task, we use a hybrid BiLSTM-CRF model to detect three types of entities: Person(PER), Location(LOC) and Organization (ORG) in Chinese financial news. The texts are crawled from multiple financial news websites, including 3000 news articles (30,000 sentences in total). All the entities are manually labeled by four graduate students major in linguistics. As financial news usually involves names of companies, stocks and official agencies, we label these names as ORG in the dataset.

4.2 Sentiment Classification

Convolutional neural networks (CNNs) are effective models for sentiment and text classification. Based on Kim's [15] work, we train a simple but effective CNN model for binary sentiment classification (positive and negative). The dataset contains 40,000 reviews collected from `https://book.douban.com/`. Each review has a star tag rated by users from one star to five stars. It could be used to build a two-class (positive/negative) [4] classification task.

5 Experiments

Table 1. Statistics of the Financial NER dataset.

	PER	LOC	ORG	Total
Training	11488	15910	29192	56590
Test	2432	3059	5874	11365
Total	13920	18969	35066	67955
Test/Total	0.1747	0.1613	0.1675	0.1672

[4] We identify one-star and two-star reviews as negative, four-star and five-star reviews as positive. Reviews with three-star are regarded as neutral comments and thus not considered.

5.1 Datasets

For intrinsic evaluation, the word similarity dataset includes 500 word pairs covering 716 unique Chinese words, which is a relatively small dataset. CA8, the word analogy dataset including 17,813 questions is a big and balanced dataset for analogical reasoning.

For the NER task, We divide the dataset into training set (25000 sentences) and test set (5000 sentences). During the training of RNN, the model will be automatically validated from the training set, so there is no validation set. Table 1 shows the distribution of three types of entities in the datasets.

The data for Sentiment Classification is divided into following three sets: 10% of the 40,000 short texts are used for test, 85% for training, the remaining for validation.

5.2 Pre-trained Word Vectors

We train word embeddings with SGNS (Skip-gram with negative-Sampling) model implemented by ngram2vec toolkit[5]. Table 2 shows the hyper-parameter settings. As shown in Table 3, six large-scale corpora ranging from 1GB to over 6 GB are used during training, including Chinese Wikipedia, Baidu-baike (an online Chinese encyclopedia), Zhihu (Chinese social QA data), People's Daily news, Sogou News and Financial News. Embedding is also trained after combining the above six corpora.

Like (Li et al., 2018)'s [19] work , while training embeddings based on each corpus, we consider integrating the n-gram and characters features, which are proved effective in training word representations [29]. Specifically, we use word bigram for n-gram features, character unigram and bigram for character features. As a result, we obtain 21 embeddings for experiments.

Table 2. Hyper-parameter settings for training word embeddings.

Window	Iteration	Dimension	Subsampling	Low-frequency threshold	Context distribution smoothing	Negative (SGNS)
5	5	300	1e-5	10	0.75	5

Table 3. Seven corpora used for training word embeddings.

	Wikipedia_zh	Zhihu	Sogou News	People's Daily	Baidu-baike	Financial	Combination
Size	1.3G	2.1G	3.7G	3.9G	4.1G	6.2G	21.3G
Token	223M	384M	649M	668M	745M	1055M	4037M
Vocab size	2129K	1117K	1226K	1664K	5422K	2785K	10653K

[5] https://github.com/zhezhaoa/ngram2vec

5.3 Results and Analysis of NER

The left part of Table 5 shows the NER results of different embeddings in ascending order of corpora sizes. We will make analysis of the results from three aspects: context features, size and domain of corpus.

Context features. As shown in table 5, the introduction of bigram and character features has brought constant improvement of performance in most scenarios. Besides, bigram features show a more distinct advantage because after integrating it, the F1 score increases in all the cases.

Corpus size. We can see that the embedding trained with the largest combination corpus always performs best, and best F1 scores in last four groups (from People's Daily to Combination) are increasing continuously with the growing size of corpora.

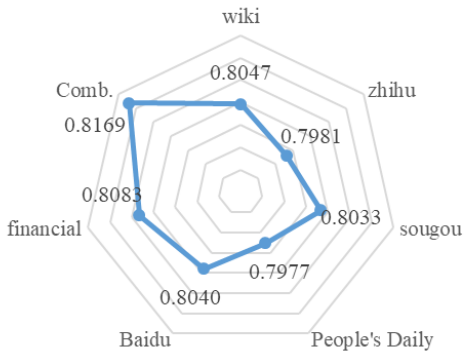

Fig. 1. Performance of different embeddings in NER task, with the best F1 score of each corpus.

Corpus domain. If we ignore the results of the combination corpus, the performance of financial embedding achieves the best among all the groups. We speculate that the reason is not only about its size, but also its domain. As the NER dataset is constructed from financial news, the embedding trained with financial domain data should have direct and positive impacts on the recognition results. Figure 1 clearly indicates the contributions of various domains.

In order to further testify the impact of corpus size and domain, we randomly sample two smaller financial corpora from the original one, and re-evaluate their performances. One of the samples is 1.3 GB, as same as the Wikipedia corpus, because we find although Wikipedia is a much smaller corpus than financial news, but its embeddings achieve comparable results with financial data.

As shown in table 4, financial embeddings of $word$ and $word + bigram$ features always outperform Wikipedia embeddings even when the size of financial corpus decreases to the same with Wikipedia (1.3 GB). The experimental results prove that both the size and domain of a corpus have important influences on embeddings.

Table 4. Comparison between wikipedia and different sizes of financial embeddings based on NER F1 scores.

	word	word+bigram	word+char
Wiki	0.7955	0.7965	0.8047
Fin._1.3G	0.7977	0.8008	0.8035
Fin._3.5G	0.7991	0.8066	0.8024
Fin._6.2G	0.8023	0.8081	0.8083

5.4 Results and Analysis of Sentiment Classification

The right part of Table 5 shows the results of sentiment classification on Book Review dataset. To keep consistent with NER task, we will also discuss the results from three aspects.

Context features. It can be seen that character and bigram features are both advantageous to model performance. More importantly, most embeddings integrated with bigram features perform the best among three kinds of features, which is consistent with NER task. Thus these two extrinsic tasks both favors more of bigram features than character features.

Corpus size. It is quite obvious that corpus size plays an important role in embedding performance. Firstly, Wikipedia and Baidu-baike are both online encyclopedia data, and with a bigger size, Baidu-baike gains much better performance than wikipedia, especially on $word$ and $word+bigram$ settings. Secondly, similar to NER task, embeddings trained with the combination corpus achieve the highest F1 scores and accuracies, indicating the size of corpus has direct and important impacts on the performance.

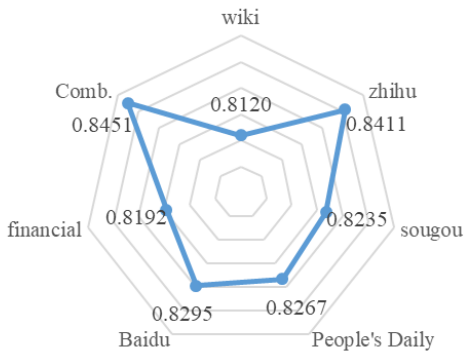

Fig. 2. Performance of different embeddings on SC task, with the best F1 score of each corpus.

Corpus domain. It can be seen that Zhihu data has a clear advantage on sentiment classification, which means the corpus domain may play more important roles on the

Table 5. Results of Named Entity Recognition and Sentiment Classification.

		Name Entity Recognition			Sentiment Classification			
		P	**R**	**F1**	**P**	**R**	**F1**	**Accuracy**
Wiki	word	0.8194	0.7730	0.7955	0.7940	0.7883	0.7851	0.7858
	word+bigram	0.8088	0.7845	0.7965	0.7829	0.7773	0.7742	0.7749
	word+char	0.8323	0.7788	**0.8047**	0.8143	0.8133	**0.8120**	**0.8121**
Zhihu	word	0.8167	0.7722	0.7938	0.8395	0.8359	0.8337	0.8339
	word+bigram	0.8120	0.7815	0.7964	0.8416	0.8409	**0.8411**	**0.8414**
	word+char	0.8287	0.7697	**0.7981**	0.8336	0.8329	0.8317	0.8317
Sogou	word	0.8306	0.7742	0.8014	0.8178	0.8176	0.8167	0.8167
	word+bigram	0.8356	0.7726	0.8028	0.8260	0.8230	**0.8235**	**0.8242**
	word+char	0.8338	0.7750	**0.8033**	0.8216	0.8219	0.8217	0.8217
People's daily	word	0.8267	0.7700	0.7974	0.8278	0.8267	0.8254	0.8255
	word+bigram	0.8192	0.7773	**0.7977**	0.8274	0.8274	**0.8267**	**0.8267**
	word+char	0.8311	0.7612	0.7946	0.8240	0.8240	0.8233	0.8233
Baidu-baike	word	0.8335	0.7714	0.8013	0.8288	0.8279	0.8267	0.8267
	word+bigram	0.8216	0.7872	**0.8040**	0.8275	0.8274	0.8275	0.8277
	word+char	0.8273	0.7691	0.7972	0.8308	0.8305	**0.8295**	**0.8295**
Financial	word	0.8344	0.7727	0.8023	0.8152	0.8137	0.8140	0.8145
	word+bigram	0.8260	0.7910	0.8081	0.8192	0.8195	**0.8192**	**0.8192**
	word+char	0.8511	0.7697	**0.8083**	0.8152	0.8147	0.8136	0.8136
Comb.	word	0.8383	0.7795	0.8078	0.8474	0.8462	0.8448	0.8448
	word+bigram	0.8374	0.7973	**0.8169**	0.8459	0.8459	**0.8451**	**0.8451**
	word+char	0.8433	0.7851	0.8131	0.8400	0.8401	0.8400	0.8402

performance of this task than that of NER. This conclusion can be reflected in at least two comparisons: (1) although Zhihu is the second smallest corpus among the corpora, the embeddings trained with zhihu data achieve significant improvements over other embeddings, e.g. nearly 7% higher than Wikipedia, and 2%-3% higher than the other news or encyclopedia corpora which are much larger than Zhihu. (2) The combination corpus is 10 times larger than Zhihu, but their results are almost the same (best F1: 0.8451 vs 0.8411, best accuracy: 0.8451 vs 0.8414).

Figure 2 clearly shows the contribution of Zhihu. One possible reason is Zhihu data is collected from a social QA website, and the Book Review data is crawled from Douban, which is also a social networking service website. Their text domains are highly similar, thus the embeddings of zhihu can make a great contribution to the classification task.

In this section, we discuss the performances of two extrinsic tasks, and find the impacts of embeddings on the tasks are not the same. More specifically, in the three kinds of context features, embeddings with $word + bigram$ features perform best in both tasks. As for sizes of corpora, larger size of some corpora can improve the performance to some extent, and the largest combination corpus achieves best results in both two tasks. However, performance does not always improve with the growing size of corpora. As for domains of corpora, domain-specific corpora do have positive and significant impacts on the performance (Financial in NER and Zhihu in sentiment classification). The influence of the domain is even more important than that of size.

5.5 Results of Intrinsic Evaluation

Table 6 shows the performance of intrinsic evaluation, including analogical reasoning on morphological and semantic relations, and word similarity. It can be clearly seen that the introduction of bigram and character features brings significant and consistent improvements on all the categories of embeddings. Furthermore, character features are especially advantageous for reasoning of morphological relations. This is because word in Chinese is composed of graphical characters, known as Hanzi, which has direct influence on Chinese morphology. Thus the introduction of character features can greatly improve the performance of morphological reasoning.

Table 6. Results of intrinsic evaluation. Mor. and Sem. belong to analogical reasoning and Sim. refers to similarity.

		Wiki	Zhihu	Sogou	People's daily	Baidu-baike	Financial	Comb.
Mor.	word	0.114	0.161	0.098	0.194	0.203	0.049	0.285
	word+bigram	0.148	0.191	0.098	0.228	0.241	0.077	0.33
	word+char	**0.395**	**0.499**	**0.343**	**0.477**	**0.417**	**0.323**	**0.543**
Sem.	word	0.188	0.156	0.239	0.406	0.319	0.225	0.489
	word+bigram	0.195	0.157	0.246	**0.407**	0.325	**0.243**	**0.492**
	word+char	**0.238**	**0.173**	**0.249**	0.403	**0.412**	0.237	0.412
Sim.	word	0.388	0.476	0.472	0.461	0.462	0.354	0.503
	word+bigram	0.397	**0.489**	**0.48**	**0.477**	**0.465**	0.350	**0.519**
	word+char	**0.414**	0.438	0.468	0.469	0.407	**0.356**	0.500

As for the size of corpus, it has direct impacts on the performance, for example, Baidu-baike outperforms Wikipedia in all the evaluation measures. Corpus domain is also an important factor in intrinsic tasks. For example, vectors trained on news data (e.g. People's Daily) are beneficial to semantic reasoning, because CA8 incorporates a lot of geography questions, and the names of countries and cities have high frequencies in news data. With the largest size and varied domains, the Combination corpus performs much better than others in both analogy and similarity tasks.

5.6 Correlation between Intrinsic Evaluation and Extrinsic Evaluation

Table 7. Correlations between evaluations. Ana. here refers to the average scores of Mor. and Sem. The strength of the correlation according to Evans (1996) [9] is : 0.00-0.19 "very weak", 0.20-0.39 "weak", 0.40-0.59 "moderate", 0.60-0.79 "strong", 0.80-1.0 "very strong".

	Inside Intrinsic Evaluation				**Between Intrinsic and Extrinsic Evaluation**			
	Mor. vs Sem.	**Mor. vs Sim.**	**Sem. vs Sim.**	**Ana. vs Sim.**	**NER vs. Ana.**	**NER vs. Sim.**	**SC vs. Ana.**	**SC vs. Sim.**
word	0.7572	0.7631	0.4931	0.6502	0.5493	0.2107	0.6464	0.7402
+bigram	0.7699	0.7012	0.4791	0.6118	0.5510	0.1534	0.4759	0.6589
+char	0.4963	0.5891	0.4317	0.5865	**-0.0658**	0.0434	0.7373	0.6456

Firstly, we can observe a lot of consistencies between intrinsic and extrinsic evaluation from above experiments.

- By introducing the character and ngram features, performances of both intrinsic and extrinsic tasks improve, but we can observe that character features are more favorable to intrinsic tasks, while ngram features prove to be more advantageous for extrinsic tasks.
- By comparing embeddings trained with corpora of different sizes and domains, we can find that larger size or similar domain can be important advantages for both intrinsic and extrinsic tasks. And the combination corpus with largest size and varied domains always performs the best.

To evaluate the correlation between these tasks objectively, we compute the correlations between above tasks by using Pearson correlation coefficient (*p*). To be specific, we compute not only the correlation between intrinsic and extrinsic, but also the correlation between two intrinsic tasks. We extract the F1 scores of three context features respectively in each task, and compute the coefficients between them.

From the results shown in Table 7, we can observe that in intrinsic evaluations, there is a consistent positive correlation between results of word analogy and word similarity in all the three types of context features. It is not surprising that the correlation between morphological reasoning and semantic reasoning is high, because they are both word analogy tasks. An interesting result is morphological reasoning has a higher correlation with word similarity than semantic reasoning. It is probably because both of the tasks involve word pairs that have same character morphemes.

Regarding the correlation between intrinsic and extrinsic evaluation, most coefficients show positive correlations, indicating intrinsic task can be good indicators of downstream tasks. The only exception is the $word + char$ embeddings in NER task. Generally, correlations between Sentiment classification task and intrinsic tasks are stronger than those between NER and intrinsic tasks. The main reason is we use a domain-specific NER dataset for test, the performance of which is largely affected by the domain issue.

Based on above analysis, we reached a couple of interesting and useful findings for evaluation of Chinese word embeddings. Firstly, intrinsic measures are useful in predicting the performances of embeddings in downstream tasks to some extent. Secondly, each task has its favorable features. We would suggest to train word embeddings with corpus that has a similar domain with the dataset. For the same domain of corpus, the bigger, the better. Moreover, extrinsic tasks favor ngram features, while intrinsic tasks favor character features. Thus it is recommended to choose suitable embeddings for each task.

6 Conclusion

This paper conducts a comprehensive study on the correlation between intrinsic and extrinsic evaluation for word embeddings. 21 word embeddings with different corpora and context features are trained and evaluated in 4 tasks: analogy reasoning and word

similarity for intrinsic evaluation, NER and Sentiment Classification for extrinsic evaluation. Experimental results prove that intrinsic and extrinsic evaluations are consistent in most cases.

Also, our study sheds some lights on how to select suitable embeddings for NLP tasks: (1) Context features can be integrated to improve the performance, and most extrinsic tasks favor ngram features, while intrinsic tasks favor character features. (2) Training Corpus is very important for the performance of word embeddings. The relevant domain is more important than size factor, especially for extrinsic tasks.

Overall, this paper presents some interesting findings for embedding evaluation, as well as several datasets which could serve as benchmarks for Chinese NLP communities. We also plan to investigate more factors that may affect the embedding performance such as different models and hyper-parameters, and to explore other downstream tasks, e.g. POS tagging and parsing.

Acknowledgements

This work is supported by the Fundamental Research Funds for the Central Universities, China Postdoctoral Science Foundation funded project (No. 2018M630095) and National Language Committee Research Program of China (No. ZDI135-42).

References

1. Bakarov, A.: A survey of word embeddings evaluation methods. arXiv preprint arXiv:1801.09536 (2018)
2. Bansal, M., Gimpel, K., Livescu, K.: Tailoring continuous word representations for dependency parsing. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). vol. 2, pp. 809–815 (2014)
3. Batchkarov, M., Kober, T., Reffin, J., Weeds, J., Weir, D.: A critique of word similarity as a method for evaluating distributional semantic models. In: Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP. pp. 7–12 (2016)
4. Camacho-Collados, J., Pilehvar, M.T., Collier, N., Navigli, R.: Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 15–26 (2017)
5. Chen, C.Y., Ma, W.Y.: Word embedding evaluation datasets and wikipedia title embedding for chinese. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). pp. 825–831 (2018)
6. Chen, X., Xu, L., Liu, Z., Sun, M., Luan, H.B.: Joint learning of character and word embeddings. In: IJCAI. pp. 1236–1242 (2015)
7. Chiu, B., Korhonen, A., Pyysalo, S.: Intrinsic evaluation of word vectors fails to predict extrinsic performance. In: Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP. pp. 1–6 (2016)
8. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. Journal of Machine Learning Research **12**(Aug), 2493–2537 (2011)
9. Evans, J.D.: Straightforward statistics for the behavioral sciences. Brooks/Cole (1996)
10. Faruqui, M., Tsvetkov, Y., Rastogi, P., Dyer, C.: Problems with evaluation of word embeddings using word similarity tasks. In: Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP. pp. 30–35 (2016)

11. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E.: Placing search in context: The concept revisited. In: Proceedings of the 10th international conference on World Wide Web. pp. 406–414. ACM (2001)
12. Gao, B., Bian, J., Liu, T.Y.: Wordrep: A benchmark for research on learning word representations. arXiv preprint arXiv:1407.1640 (2014)
13. Ghannay, S., Favre, B., Esteve, Y., Camelin, N.: Word embedding evaluation and combination. In: LREC. pp. 300–305 (2016)
14. Gurnani, N.: Hypothesis testing based intrinsic evaluation of word embeddings. arXiv preprint arXiv:1709.00831 (2017)
15. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
16. Levy, O., Goldberg, Y.: Dependency-based word embeddings. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). vol. 2, pp. 302–308 (2014)
17. Levy, O., Goldberg, Y.: Linguistic regularities in sparse and explicit word representations. In: Proceedings of the eighteenth conference on computational natural language learning. pp. 171–180 (2014)
18. Levy, O., Goldberg, Y., Dagan, I.: Improving distributional similarity with lessons learned from word embeddings. Transactions of the Association for Computational Linguistics **3**, 211–225 (2015)
19. Li, S., Zhao, Z., Hu, R., Li, W., Liu, T., Du, X.: Analogical reasoning on chinese morphological and semantic relations. arXiv preprint arXiv:1805.06504 (2018)
20. Li, Y., Li, W., Sun, F., Li, S.: Component-enhanced chinese character embeddings. arXiv preprint arXiv:1508.06669 (2015)
21. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
22. Nayak, N., Angeli, G., Manning, C.D.: Evaluating word embeddings using a representative suite of practical tasks. In: Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP. pp. 19–23 (2016)
23. Schnabel, T., Labutov, I., Mimno, D., Joachims, T.: Evaluation methods for unsupervised word embeddings. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 298–307 (2015)
24. Tjong Kim Sang, E.F.: Introduction to the conll-2002 shared task: Language-independent named entity recognition. In: Proceedings of CoNLL-2002. pp. 155–158. Taipei, Taiwan (2002)
25. Turian, J., Ratinov, L., Bengio, Y.: Word representations: a simple and general method for semi-supervised learning. In: Proceedings of the 48th annual meeting of the association for computational linguistics. pp. 384–394. Association for Computational Linguistics (2010)
26. Wu, Y., Li, W.: Overview of the nlpcc-iccpol 2016 shared task: chinese word similarity measurement. In: Natural Language Understanding and Intelligent Applications, pp. 828–839. Springer (2016)
27. Xu, J., Liu, J., Zhang, L., Li, Z., Chen, H.: Improve chinese word embeddings by exploiting internal structure. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1041–1050 (2016)
28. Zhai, M., Tan, J., Choi, J.D.: Intrinsic and extrinsic evaluations of word embeddings. In: AAAI. pp. 4282–4283 (2016)
29. Zhao, Z., Liu, T., Li, S., Li, B., Du, X.: Ngram2vec: Learning improved word representations from ngram co-occurrence statistics. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 244–253 (2017)