Chinese GloVe Word Vector Performance on CA8

Yiran Mao
McGill University
yiran.mao@mail.mcgill.ca

Silan He
McGill University
silan.he@mail.mcgill.ca

Abstract

Recent methods for learning vector space representations of words have succeeded in capturing fine-grained semantic and syntactic regularities in the English language. The result is the new global log-bilinear regression model GloVe (Jeffrey Pennington, Richard Socher, and Christopher D. Manning, 2014) that combines the advantages of the two major model families in the literature: global matrix factorization and local context window methods.

Similar strides are being made in China, specifically on Chinese morphological and semantic relations. We analyze the effectiveness of GloVe trained on a standardized Chinese corpus versus the performance of current Chinese word vector models (Li et al., 2018a) trained on the same corpus. The Chinese GloVe vector is found to be marginally better semantically while the native Chinese vectors as described by Shen Li et al. performs much better morphologically. Positive Pointwise Mutual Information outclasses the performance of GloVe in terms of semantics while being slightly better morphologically.

1 Introduction

Semantic vector space models of language represent each word with a real-valued vector. Given the word representations, analogy questions can be automatically solved via vector computation.

For example, "apples - apple + car = cars" for morphological regularities and "king - man + woman = queen" for semantic regularities (Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, 2013). These vectors can also be used in a variety of applications such as:

- document classification (Fabrizio Sebastiani, 2002)
- question answering (Stefanie Tellex, Boris Katz, Jimmy Lin, Aaron Fernandes, and Gregory Marton, 2003)
- translating unknown words (Philippe Langlais and Alexandre Patry, 2007)
- detecting semantic relations (Amac Herdagdelen and Marco Baroni, 2009)
- inducing morphological transformations (Radu Soricut and Franz Josef Och, 2015)

However, few attempts have been made in specifically Chinese analogical reasoning. It is well known that linguistic regularities vary a lot among different languages. For example, Chinese is a typical analytic language which lacks inflection.

In this report, we analyze the model properties of GloVe compared to those of existing Chinese word vector representations (Li et al., 2018a). We further compare performance by subjecting all models to the new CA8 standard benchmark for evaluation of Chinese word embedding as presented by Shen Li et al.

2 Word Vectors/Embeddings

2.1 GloVe

GloVe trains on unigram word ratios of co-occurence probabilities than on the word probabli-

ties themselves. The ratio is shown to be able to better distinguish relevant words from irrelevant words than the word probabilities themselves. The ratio is then isolated as a least squares problem that is weighted by a function between 0 and 1.

2.2 Word2vec / Skip-Gram with Negative Sampling (SGNS)

Skip-Gram consists of a neural network with a small hidden layer. Using the pairs of co-occuring words and words/ngrams as input, learning the weights of the hidden layer will give us the word vectors that we are looking for (Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, 2013).

Negative Sampling is a process in which you only update a small percentage of weights as opposed to all the weights in the hidden layer (Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean, 2013).

Currently, SGNS is more versatile than GloVe as it samples word co-occurence context features in a variety of ways. We believe this will influence the morphological performance of GloVe compared to the various SGNS.

2.3 Pointwise Positive Mutual Information

It turns out, however, that simple frequency isn't the best measure of association between words. One problem is that raw frequency is very skewed and not very discriminative. If we want to know what kinds of contexts are shared by apricot and pineapple but not by digital and information, we're not going to get good discrimination from words like 'the', 'it', or 'they', which occur frequently with all sorts of words and aren't informative about any particular word (Daniel Jurafsky and James H. Martin, 2000).

Therefore, sometimes we replace this raw frequency with positive pointwise mutual information:

$$PPMI(w,c) = max(log\_2 \frac{P(w,c)}{P(w)P(c)}) \quad (1)$$

where w refers to word in question and c refers to context words.

Good collocation pairs will have high PMI or pointwise mutual information because the probability of co-occurrence is only slightly lower than the probabilities of occurrence of each word. Conversely, a pair of words whose probabilities of occurrence are considerably higher than their probability of co-occurrence gets a small PMI score.

3 Linguistic Regularities Chinese

There was a belief that Chinese is a morphologically impoverished language since a morpheme mostly corresponds to an orthographic character, and it lacks apparent distinctions between roots and affixes. However, it was determined that Chinese merely operates on a different morphological system (Jerome L Packard, 2000).

Li et al. clarifies this special system by mapping its morphological analogies into two processes: reduplication and semi-affixation. Both processes are tested in the CA8.

3.1 Reduplication

In absence of inflection, the Chinese language has a mechanism called reduplication in which nouns, adjectives,verbs, adverbs and measure words can be repeated to generate a new meaning. This is coined as reduplication.

```
bà(dad) -> bà-bà(dad)
tiān(day) -> tiān-tiān(everyday)
shuō(say) -> shuō-shuo(say a little)
kàn(look)  kàn-kàn(have a brief look)
dà(big)  dà-dà(very big; greatly)
```

3.2 Semi-affixation

Affixation is a morphological process whereby a bound morpheme (an affix) is attached to roots or stems to form new language units. While Chinese is mostly absent of affixes, Chinese has been shown to have similar construct that in itself is a word but can also be used to form new words when attached to other words. They are dubbed semi-affixes (Yuehua Liu, Wenyu Pan, and Wei Gu, 2001).

```
yī(one)  dì-yī(first)
èr(two)  dì-èr(second)
pàng(fat)  pàng-zi(a fat man)
shòu(thin)  shòu-zi(a thin man)
```

For example, the semi-prefix "dì-" could be added to numerals to form ordinal numbers, and the semi-suffix "-zi" can transform an adjective into a noun.

4 Experiments

4.1 Preprocessing and Corpus

All word embeddings were trained on data from WeiBo, a chinese microblog, provided by NLPIR Lab (Lab, 2018). Though there were many other options available, we were unable to find the exact corpus that Li et al. used for the other ones, so we settled on a publicly available corpus to train on.

All the text data used in our experiments are preprocessed via the following steps:

- Remove the html and xml tags from the texts and set the encoding as utf-8. Digits and punctuation in the exchanges are kept.
- Conduct Chinese word segmentation with Jieba (Inc., 2018)

Refer to Appendix A for an example.

4.2 GloVe Chinese Vector

Following the preprocessing, all the text data was plugged into the GloVe code in order to generate the word vector representation. The code is available on the GloVe github (stanfordnlp, 2014). Simply run the following files sequentially:

- vocab_count.c
- cooccur.c
- shuffle.c
- glove.c

Set dimension to 300 because it is the standard that the SGNS are trained (Li et al., 2018a).

4.3 Word2Vec, Skip-Gram with Negative Sampling (SGNS)

These all consist of dense dim=300 vectors trained on the same Weibo corpus (Li et al., 2018b). We analyze all the vectors trained with different context features: word, word + ngram, word + character, word + character + ngram.

4.4 Positive Pointwise Mutual Information (PPMI)

This consists of a sparse vector trained on the same Weibo corpus (Li et al., 2018b). It is expected that this will perform much better semantically as it performs better semantically than SGNS. We were only able to run the test for the vector trained with unigrams as context features.

4.5 Evaluation method: CA8

Though Ca_translated had been widely used in the evaluations of word embeddings, it should not serve as a reliable benchmark since it only includes 134 unique Chinese words in three semantic relations (capital, state, and family). In addition, morphological knowledge is not even considered in the dataset.

Therefore, we evaluate the quality of the GloVe word vectors and the pre-trained Word2Vec word vectors using the specially designed CA8 evaluation dataset. The CA8 dataset includes explicitly chosen analogical reasoning tasks designed to accurately evaluate both morphological analogies and semantic knowledge reasoning. The evaluation measure used is a simple accuracy calculation.

The semantics test splits the dataset into 4 categories: geography, nature, history and people. This allows users of CA8 to pinpoint the weaknesses of their word vectors.

The morphological tests are also separated into 4 categories: reduplication A, prefix, reduplication AB and suffix. Each contain tests pertaining to the described construct.

For each category, the accuracy of two analogy test dubbed additive (add) and multiplicative (mul) are computed. Following are the formulas describing them:

$mul_sim = sim_b * sim_c * np.reciprocal(sim_a + 0.01)$

$add_sim = -sim_a + sim_b + sim_c$

5 Results

5.1 Semantics test

On the CA8 dataset, we observe that GloVe actually performs better than SGNS in semantic relations. GloVe's weighting function must be slightly more sensitive to infrequent and specific word pairs, which are beneficial for semantic relations. GloVe

word vectors	glove	word	word + ngram	word + character	word + character + ngram	ppmi
geography	0.301/0.296	0.259/0.266	0.291/0.303	0.236/0.238	0.236/0.241	**0.364/0.497**
nature	0.176/0.158	0.153/0.154	0.173/0.160	0.168/0.16	0.196/0.191	**0.189/0.209**
history	0.033/0.015	0.036/0.036	0.031/0.031	0.02/0.026	0.015/0.02	**0.122/0.056**
people	0.171/0.152	0.092/0.088	0.126/0.118	0.084/0.084	0.097/0.105	**0.179/0.223**
total	0.227/0.215	0.185/0.188	0.214/0.215	0.175/0.175	0.185/0.188	**0.269/0.346**

Table 1: Semantics accuracy test results. (add/mul accuracy)

word vectors	glove	word	word + ngram	word + character	word + character + ngram	ppmi
reduplication A	0.044/0.046	0.060/0.061	0.066/0.076	0.189/0/187	**0.235/0.255**	0.017/0.042
prefix	0.028/0.024	0.063/0.063	0.070/0.073	0.359/0.366	**0.371/0.372**	0.043/0.052
reduplication AB	0.028/0.019	0.050/0.050	0.073/0.076	**0.493/0.471**	0.494/0.464	0.025/0.045
suffix	0.054/0.049	0.089/0.089	0.105/0.110	**0.535/0.528**	0.525/0.524	0.071/0.116
total	0.039/0.035	0.066/0.065	0.079/0.084	0.388/0.383	**0.406/0.4**	0.039/0.064

Table 2: Morphological accuracy test results. (add/mul accuracy)

performs worse than PPMI in terms of semantics. PPMI's scoring make it much more sensitive to infrequent and specific word pairs than GloVe.

5.2 Morphology test

We can observe that on CA8 dataset, skip-gram model with negative sampling (word2vec) representations perform better in analogical reasoning of morphological relations. This is probably because the reasoning on morphological relations relies more on common words in context. In addition, the training procedure of word2vec favors frequent word pairs.

SGNS performs much better morphologically than PPMI. By extension, GloVe was actually worse morphologically than ppmi.

GloVe being trained on unigrams co-occurences was especially weak in this regard for the same reason.

6 Conclusion

Recently, a lot of progress has been made with regards to word embeddings/vectors for the Chinese language. These word embeddings can be used in various downstream natural language processing tasks.

GloVe is a better and faster word2vec or SGNS semantically (Jeffrey Pennington, Richard Socher, and Christopher D. Manning, 2014). However, SGNS only truly shines in the chinese morphology tests and it is able to holds its own semantically to GloVe depending on the contextual features used. In addition, PPMI is much better than GloVe semantically according to CA8.

For future work, using word + ngram, word + character, word + character + ngram context features into GloVe could be very interesting to see if that's enough to improve GloVe for morphological reasoning in Chinese.

References

Amac Herdagdelen and Marco Baroni. 2009. Bagpack: A general framework to represent semantic relations. page 33–40.

Daniel Jurafsky and James H. Martin. 2000. *Speech and Language Processing*.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*.

Baidu Inc. 2018. "jieba" (chinese for "to stutter") chinese text segmentation: built to be the best python chinese word segmentation module.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation.

Jerome L Packard. 2000. The morphology of chinese: A linguistic and cognitive approach.

NLPIR Lab. 2018. Weibo microblog data.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018a. Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143. Association for Computational Linguistics.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018b. Chinese word vectors .

Philippe Langlais and Alexandre Patry. 2007. Translating unknown words by analogical learning. *EMNLP-CoNLL*, page 877–886.

Radu Soricut and Franz Josef Och. 2015. Unsupervised morphology induction using word embeddings. page 1627–1637.

stanfordnlp. 2014. Glove: Global vectors for word representation.

Stefanie Tellex, Boris Katz, Jimmy Lin, Aaron Fernandes, and Gregory Marton. 2003. Quantitative evaluation of passage retrieval algorithms for question answering. *Proceedings of the SIGIR Conference on Research and Development in Informaion Retrieval*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *ICLR Work-shop Papers*.

Yuehua Liu, Wenyu Pan, and Wei Gu. 2001. Practical grammar of modern chinese. *The Commercial Press*.

A Appendix A: Weibo Data Preprocessing Example

(Could not find a latex package that was able to capture all the chinese characters in teh excerpt. Had to resort to pictures hence the terrible formatting.)

For example, following is a excerpt of raw Weibo data.

```
2815    2               0
        Sat Apr 12 04:12:49
+0800 2014
        3698385041656369
        3698385041656369

        3698385041656369
        0
        0       <a
href="http://app.weibo.com/t/
feed/9ksdit"
rel="nofollow">iPhone客户端</a>
        回复@停转的牧马:Me
too! :)//@停转的牧马:Hope you
come to china[害羞]//@CodyKarey:
[哈哈][嘻嘻] //@朱皓轩
__Wanting: //@张峡浩SeannyLuck:黑
乎乎，我喜欢右下角上边那种 //@曲婉婷
Wanting:非常怀念北美巡演。
@CodyKarey 快来看，还有你呢！你也可以
这么可爱！[哈哈]
        5060511691      00
        3695045658088034
        1397717462226
```

The next excerpt has all the html and xml tags removed. Punctuation is kept.

```
回复@停转的牧马 : ! : ) / /@停转的牧
马: [害羞] / / @ : [哈哈] [嘻
嘻] / / @朱皓轩 __ : / / @张峡浩:黑
乎乎 ,我喜欢右下角上边那种/ /@ 曲婉婷:
非常怀念北美巡演。@ 快来看,还有你呢 !
你也可以这么可爱![ 哈哈]
```

Finally, we apply Jieba.

```
回复 @ 停转 的 牧马 : ! : ) / / @
停转 的 牧马 : [ 害羞 ] / / @ :
[ 哈哈 ] [ 嘻嘻 ] / / @ 朱皓轩
__ : / / @ 张峡浩 : 黑乎乎 ， 我 喜
欢 右下角 上边 那种 / / @ 曲婉婷 :
非常 怀念 北美 巡演 。 @ 快 来看 ，
还有 你 呢 ！ 你 也 可以 这么 可爱 !
[ 哈哈 ]
```