Traversal-Free Word Vector Evaluation in Analogy Space

Xiaoyin Che, Nico Ring, Willi Raschkowski, Haojin Yang, Christoph Meinel
Hasso Plattner Institute, University of Potsdam
Campus Griebnitzsee, D-14440 Potsdam, Germany
{xiaoyin.che, haojin.yang, christoph.meinel}@hpi.de
{nico.ring, willi.raschkowskil}@student.hpi.uni-potsdam.de

Abstract

In this paper, we propose an alternative evaluating metric for word analogy questions (*A to B is as C to D*) in word vector evaluation. Different from the traditional method which predicts the fourth word by the given three, we measure the similarity directly on the "relations" of two pairs of given words, just as shifting the relation vectors into a new analogy space. Cosine and Euclidean distances are then calculated as measurements. Observation and experiments shows the proposed analogy space evaluation could offer a more comprehensive evaluating result on word vectors with word analogy questions. Meanwhile, computational complexity are remarkably reduced by avoiding traversing the vocabulary.

1 Introduction

In recent years, word vector, or addressed as word embedding or distributed vector representation of word, achieves high popularity in NLP (***N**atural **L**anguage **P**rocessing*) applications. A word vector is a real-valued vector, which is quite low-dimensional when comparing with traditional one-hot representation of words. The theory behind is believed to be the early concept of distributional representation (Hinton, 1986), and modern word vector derives from the training process of neural language models (Bengio et al., 2003).

The usage of word vectors has been proven highly efficient and successful by various NLP tasks (Collobert et al., 2011), which further spurs the technical developments to achieve word vectors with better quality, such as Word2Vec (Mikolov et al., 2013a), GloVe (Pennington et al., 2014), Word2Vecf (Levy and Goldberg, 2014), LexVec (Salle et al., 2016), FastText (Bojanowski et al., 2016), *etc*.

However, discussion about how to evaluate the quality of word vectors remains open. Except for actual applications, most frequently used evaluation tasks are word similarity and word analogy. A word similarity task is to find the nearest word in the vector space of the given word, based on the theory that words with similar meanings should gather together. Although it is widely used, arguments are made to question its capability (Batchkarov et al., 2016; Faruqui et al., 2016).

While in a word analogy test, three words A, B and C are given and the goal is to find a fourth word D, which logically conforms "A to B is as C to D". Word analogy test has a long history of being used in examinations or IQ tests for human (McClelland, 1973; Sternberg, 1985) and is introduced into word vector evaluation by Mikolov *et al.* (2013b). After that, it has been widely applied.

Efforts are made to improve the original analogy metric, such as using PAIRDIRECTION to replace 3COSADD in calculation (Levy et al., 2014) or taking multiple word pairs into consideration (Drozd et al., 2016), but the goal is still to find word D from the vocabulary. Besides, Linzen (2016) made a thorough assessment of word analogy test, and the most prominent finding is that if not exclude three given words, the prediction of D would almost always be C (91%) or B (5%), especially when the lineal offset between words is small. This phenomenon would arouse the doubt, that whether we are searching for a word D which holds the same logic to C just as B to A, or actually searching for the nearest word of C? Furthermore, the general accuracy decline in reversed analogy also suggests the incertainty of current analogy evaluation metric.

In this paper, we would dig deeper into the limitations of current analogy evaluation metric in

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1–5
Copenhagen, Denmark, September 7–11, 2017. ©2017 Association for Computational Linguistics

Table 1: Examples of Traditional Word Analogy Evaluation Result (Words in order of A, B, C & D)

Grammar-1	knowing		knew		selling		sold	
Predictions	thought	0.573	know	0.481	purchased	0.520	**sold**	**0.568**
	know	0.504	**knew**	**0.449**	resold	0.506	sell	0.535
	wanted	0.494	Knowing	0.441	**selling**	**0.486**	bought	0.528
	knowing	0.489	figured	0.404	sale	0.484	buying	0.486
Grammar-2	**looking**		**looked**		**shrinking**		**shrank**	
Predictions	look	0.540	**looked**	**0.536**	**shrinking**	**0.560**	shrunk	0.618
	looking	**0.526**	look	0.493	unexpectedly_shrank	0.478	**shrank**	**0.589**
	looks	0.521	looks	0.415	downwardly_revised	0.468	dwindled	0.498
	seemed	0.439	expecting	0.410	contraction	0.454	shrink	0.498

Section 2 and propose our simple alternative plan in Section 3, which is called "Analogy Space Evaluation". A significant difference of our approach is that we avoid traversing vocabulary from time to time. Experiments are presented in Section 4 and finally come the conclusion and discussion.

2 Limitations of Traditional Metric

In traditional word analogy evaluation, by given word pairs (A, B) and (C, D) with same syntactic or semantic relation, the goal is to find the nearest word to "$C+B-A$" in the vector space by Cosine similarity and check whether the word obtained is D. Practically some approaches use unit vector of A, B and C in "$C + B - A$", such as widely used Word2Vec. Anyway, the return value of such a word analogy question is in Boolean type.

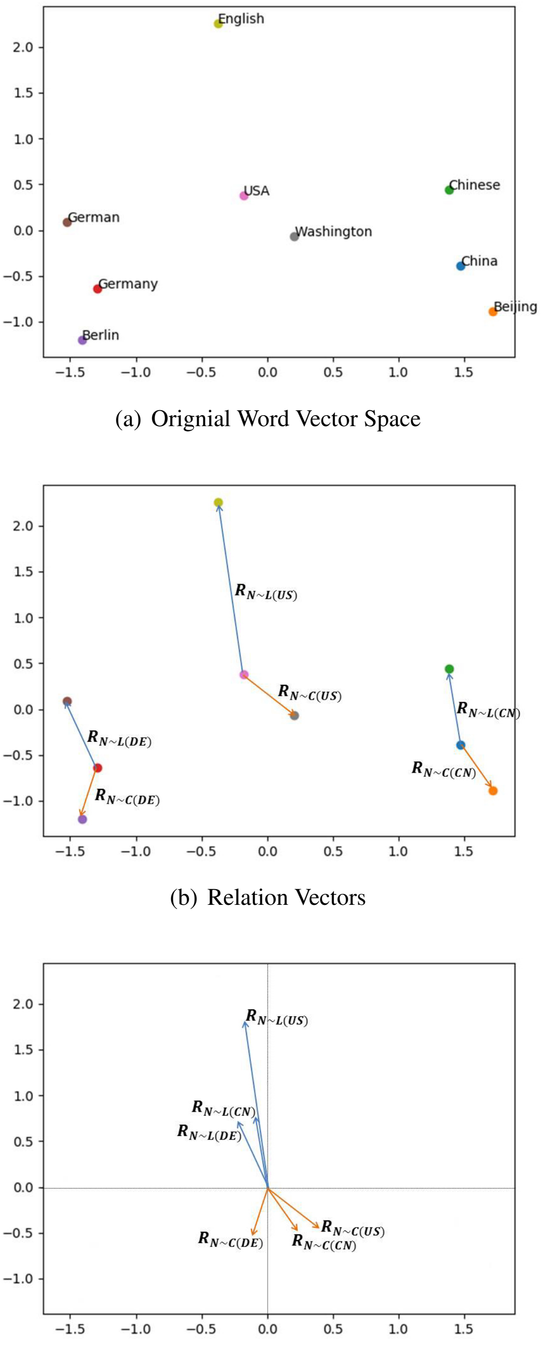
Generally, evaluating word vectors requires thousands of word analogy questions, which return thousands of Boolean values to calculate the accuracy from a macro perspective: how many supposed D have been successfully predicted. However, if we treat each question as an independent target in a micro aspect, result in Boolean type suffers an unneglectable information loss: true or false cannot quantitatively manifest the extent of how true or how false. For instance, it does not matter whether D is the 2nd nearest word to "$C + B - A$" or the 100th.

Another limitation of traditional metric is the deficiency in comprehensiveness. In a typical "A to B is as C to D" analogy, there are in fact 4 prediction choices, although in some analogies like "Nation-Currency" or "Nation-Language", available choices could drop to 2, since in reverse logic the answer is not unique. A single prediction on D is not enough to represent the quality of all 4 word vectors trained.

For better illustration, we run widely used "GoogleNews-vectors-negative300.bin" on default Word2Vec English analogy test and extract two examples to Table 1. All 4 words in example analogy questions are predicted and top 4 results are presented accordingly. From Table 1, it is clear that no matter with absolute value or average ranking of desired word in predictions, situation in Grammar-2 is apparently better than Grammar-1. However, because only word D is predicted by traditional metric, Grammar-1 would return a positive result while Grammar-2 is negative, which obviously fails to correctly represent the quality of corresponding word vectors trained.

In default Word2Vec analogy test, there is always another analogy question, which in fact predict word B of the original question. But there is no reverse logic prediction for A and C. So in final accuracy calculation, these two sets of words in Table 1 contribute the same precision of 0.5, which still cannot reflect the quality difference between these two sets of word vectors trained. Perhaps, 4 analogy questions are needed, but that would lead to another issue: higher complexity. Every time when searching for a nearest word, cosine similarity must be calculated with each word in the vocabulary. When the testing set is large, it may take quite a long time, and the time would be doubled if all 4 possible questions are included. Moreover, the majority of words in the vocabulary are actually unrelated with the prediction target. Calculating these words is simply wasting time.

Based on all above reasons, we aim to offer an alternative metric for word analogy evaluation, by constructing a new analogy space based on the relation vectors achieved from analogy questions, in order to solve existing limitations in quantification, comprehensiveness and complexity.



(a) Orignial Word Vector Space

(b) Relation Vectors

(c) Final Analogy Space

Figure 1: Analogy Space Illustration

Table 2: Analogy Space Evaluation (Micro)

Analogy	Cos.	Euc.	N-Cos.	N-Euc.
Grammar-1	0.114	0.334	0.115	0.332
Grammar-2	0.324	0.410	0.320	0.415
NC: US-CN	0.310	0.380	0.314	0.356
NC: US-DE	0.367	0.423	0.376	0.411
NC: DE-CN	0.496	0.492	0.508	0.495
NL: US-CN	0.452	0.420	0.451	0.405
NL: US-DE	0.438	0.430	0.441	0.418
NL: DE-CN	0.712	0.617	0.714	0.619

3 Analogy Space Evaluation

Proposed analogy space shares same dimensionality of original word vector space. For each analogy question, two relation vectors can be found in original word vector space, just as the definition of PAIRDIRECTION by Levy *et al.* (2014). Mathematically, the value of such a relation vector is the same as the position of the ending point if we take the starting point as the space origin. This is simply the new analogy space: shifting all relation vectors to the space origin, so each point in this new space represents a relation between a pair of words given in the analogy question. Figure 1 illustrates this process by several example words of “Nation-Capital” and “Nation-Language” analogies (*extracted from same test of Table 1, visualized by PCA*).

Naturally, we expect relations with same or similar logic gather together in the analogy space. In order to quantitatively evaluate the similarity, we prepare four different measurements, based on Cosine similarity or Euclidean distance respectively. If we denote the vectors of word A, B, C and D as $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ and $\mathbf{d}$, then

$$Cos. = \frac{(\mathbf{b}-\mathbf{a})\cdot(\mathbf{d}-\mathbf{c})}{\|\mathbf{b}-\mathbf{a}\|\|\mathbf{d}-\mathbf{c}\|} \tag{1}$$

$$Euc. = 1 - \frac{\|(\mathbf{b}-\mathbf{a})-(\mathbf{d}-\mathbf{c})\|}{\|\mathbf{b}-\mathbf{a}\|+\|\mathbf{d}-\mathbf{c}\|} \tag{2}$$

while $Cos. \in [-1,1]$ and $Euc. \in [0,1]$. *N-Cos.* and *N-Euc.* have similar definitions, but using unit word vectors in calculation. Table 2 shows the result of examples mentioned in Table 1 and Figure 1. Among them, “NC:DE-CN” and “NL:DE-CN” succeed 2/2 in traditional nearest word evaluation, while all others achieve 1/2.

It’s clear that proposed measurements could better represent the quality of these involved words or relations in a quantitative way. As already mentioned, words in Grammar-2 are considered better trained than Grammar-1, and this difference can be captured by proposed measurements only. And for NCs and NLs, traditional metric reports exactly the same accuracy, but as we can see, detailed similarities differ a lot. We believe these phenomena could help word analogy evaluation in the micro aspect.

Table 3: Analogy Space Evaluation (Macro)

WV Set	Voc.	Traditional		Proposed					SBD	
		Accu.	Time	Cos.	Euc.	N-Cos.	N-Euc.	Time	4C	2C
EN-w5-i5	1.8M	0.697	24’56”	0.325	0.419	0.325	0.420	0’38”	0.575	0.820
EN-w10-i10	1.8M	0.692	24’41”	0.314	0.414	0.315	0.415	0’37”	0.573	0.822
GoogleNews	3M	0.737	30’48”	0.352	0.431	0.352	0.431	1’09”	0.580	0.824
DE-w5-cbow	1.8M	0.465	92’39”	0.324	0.418	0.335	0.423	1’33”	–	–
DE-w5-sg	1.8M	0.434	92’09”	0.259	0.389	0.260	0.392	1’35”	–	–
DE-w10-cbow	1.8M	0.463	89’22”	0.318	0.416	0.331	0.422	1’37”	0.640	0.779
DE-w10-sg	1.8M	0.412	94’13”	0.251	0.385	0.254	0.389	1’34”	0.619	0.767

4 Macro Experiments

In this section, we would do some experiments on complete analogy question sets and discuss complexity. For English word vectors, we trained two sets on Wikipedia dump with different window size (*w*) and iteration (*i*) by Skip-Gram model, with same dimensionality of 300. They would further be compared with GoogleNews public set. We will evaluate these sets with proposed measurements, along with traditional analogy evaluation result and the performances of a downstream application: ***S**entence **B**oundary **D**etection (SBD).* Details of SBD implementation can be found in references (Che et al., 2016a,b).

Beside of English test, we also conducted several tests in German. Leipzig dataset (Goldhahn et al., 2012) are used to training German word vectors with Word2Vec toolkit. Then the vectors with different training configurations are evaluated by a set of analogy questions, which contains 2834 semantic questions in 18 categories (*including some reverse logics*) and 77886 syntactic questions in 9 categories. We have uploaded these analogy questions in German for public access[1].

Table 3 shows the results and time expenditures of these experiments. It is clear that proposed measurements have same trend with traditional metric, which means once set X achieves better result than set Y in traditional test, it would also do better in proposed alternatives. Performances in downstream application SBD are also fit this trend in general. Meanwhile, proposed evaluation could significantly save time, approximately 95%. These facts prove that we can achieve same performance within way less time.

However, we also found some limitations. The absolute difference between different vector sets in proposed measurements is smaller, which make it difficult to distinguish, especially with *Euc.* and *N-Euc.* It is also unclear that which measurement from the four proposed could be the optimized option.

5 Conclusion & Discussion

In this paper, we discuss some limitations of traditional word analogy evaluation metric in word vector evaluation, and then propose a simple alternative plan called “Analogy Space Evaluation”, which directly measures the relation vectors between given pairs of words, instead of traversing the vocabulary to seek the nearest word of the target. Experiments shows that proposed approach serves as good as traditional metric in performance, but reduces the computational complexity significantly.

This effort can be simply applied on any existing word analogy tasks. Frankly speaking, we cannot claim that our method outperforms the original, except for the complexity part. But complexity does matter. Currently analogy tasks generally contain tens of thousands questions, so traditional traversal-based evaluation can still manage. However, we would definitely want to test higher portion of words in the vocabulary, and with the efforts from the whole community, we may have a “nearly optimized” test set someday with up to million words involved. At that time, traversal-free could be a highly desirable quality.

As far as we know, there is no widely acknowledged benchmark which can be used to test new evaluation methods, so our effort remains estimation. In the future, we would attempt to implement more real applications, just as SBD mentioned in this paper, and take their performances as feedbacks, in order to contribute in this dilemma of “Evaluation of Evaluation”.

[1] https://drive.google.com/open?id=
0B13Cc1a7ebTuaE83NEtyemM4aGM

References

Miroslav Batchkarov, Thomas Kober, Jeremy Reffin, Julie Weeds, and David Weir. 2016. A critique of word similarity as a method for evaluating distributional semantic models .

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research* 3(Feb):1137–1155.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* .

Xiaoyin Che, Sheng Luo, Haojin Yang, and Christoph Meinel. 2016a. Sentence boundary detection based on parallel lexical and acoustic models. *Interspeech 2016* pages 2528–2532.

Xiaoyin Che, Cheng Wang, Haojin Yang, and Christoph Meinel. 2016b. Punctuation prediction for unsegmented transcript based on word vector. In *The 10th International Conference on Language Resources and Evaluation (LREC)*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12(Aug):2493–2537.

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. 2016. Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers,* pages 3519–3530.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. *arXiv preprint arXiv:1605.02276* .

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *LREC*. pages 759–765.

Geoffrey E Hinton. 1986. Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society*. Amherst, MA, volume 1, page 12.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Volume 2.*. The Association for Computer Linguistics.

Omer Levy, Yoav Goldberg, and Israel Ramat-Gan. 2014. Linguistic regularities in sparse and explicit word representations. In *CoNLL*. pages 171–180.

Tal Linzen. 2016. Issues in evaluating semantic spaces using word analogies. *arXiv preprint arXiv:1606.07736* .

David C McClelland. 1973. Testing for competence rather than for" intelligence.". *American psychologist* 28(1):1.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *HLT-NAACL*. pages 746–751.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)* 12:1532–1543.

Alexandre Salle, Marco Idiart, and Aline Villavicencio. 2016. Matrix factorization using window sampling and negative sampling for improved word representations. In *The 54th Annual Meeting of the Association for Computational Linguistics*. page 419.

Robert J Sternberg. 1985. *Beyond IQ: A triarchic theory of human intelligence*. CUP Archive.