The word analogy testing caveat

Natalie Schluter
Department of Computer Science
IT University of Copenhagen
Copenhagen, Denmark
natschluter@itu.dk

Abstract

There are some important problems in the evaluation of word embeddings using standard word analogy tests. In particular, in virtue of the assumptions made by systems generating the embeddings, these remain tests over randomness. We show that even supposing there were such word analogy regularities that should be detected in the word embeddings obtained via unsupervised means, standard word analogy test implementation practices provide distorted or contrived results. We raise concerns regarding the use of Principal Component Analysis to 2 or 3 dimensions as a provision of visual evidence for the existence of word analogy relations in embeddings. Finally, we propose some solutions to these problems.

1 Introduction

Continuous dense representations of words, or *word embeddings*, are d-dimensional vectors obtained from raw unannotated text. As weight vectors, they provide, given some model, predictions of either (1) some context of a word, or (2) a word given its context. The word embeddings are meant to reflect distributional structure as a proxy to semantics and syntax à la Harris (Harris, 1954). A natural and desirable effect of such context driven learning of word embeddings is distributional similarity, whereby words that are similar to each other will tend to group together in the target hyperspace. Thus *Frenchman*, *Spaniard*, and *Dane* should group together, as should *loves*, *likes*, and *admires*, or *French*, *Spanish* and *Danish*, as respectively "a set of words for humans from specific countries", "a set of present tense transitive verbs denoting fondness", and "a set of languages".

By employing a transfer learning approach with the use of word embeddings in the place of one-hot word feature vectors, word embeddings obtained in this way have been shown to both simplify and improve the performance of systems across a wide range of NLP tasks. Moreover, word embeddings trained this way and used as initial word representations are now commonly understood to improve the learning process in neural network based systems across the same array of NLP tasks.

There has been some progress in understanding why these representations work so well and a number of simple tasks developed to evaluate them independently such as (1) word similarity tests, (2) synonym selection tests, and (3) word analogy tests, in addition to a variety of possible downstream system tests. That the distributional representations of words should reflect semantic similarity (i.e., as tested by (1) and (2)) is inherent in the definition of the word embedding learning task. However that similar *relations* between words should be described by word embeddings obtained this way is not straightforward. There are also standard engineering practices in analogy evaluations that would prevent accurate analogy testing even if it were applicable.

In this paper, we hope to survey some main problems concerning the word analogy test as it is currently being calculated, in three separate directions:

1. **Theoretical assumption misalignment:** A purely distributional hypothesis misaligns with testing for analogy relations.

2. **Poor conventional engineering choices:**

   (a) Word embeddings are normalised and therefore distorted before testing.

   (b) Premise vectors are excluded before prediction.

3. **Problematic visual evidence:** Visualisations

Proceedings of NAACL-HLT 2018, pages 242–246
New Orleans, Louisiana, June 1 - 6, 2018.

based on the output of Principal Component Analysis (PCA) are misleading.

2 The word analogy tests and associated benchmarking data

The **word analogy assumption**, introduced by Mikolov et al. (2013b), elaborated with more precision by Levy and Goldberg (2014) and adapted partially from Jurgens et al. (2012) goes as follows. Suppose we have representations for two pairs of words

$$(\mathbf{a_1}, \mathbf{b_1}), (\mathbf{a_2}, \mathbf{b_2}) \quad (1)$$

having an analogous syntactic or semantic relation: $\mathbf{a_1}$ is to $\mathbf{b_1}$ what $\mathbf{a_2}$ is to $\mathbf{b_2}$. By the word analogy assumption, this analogous relation should be represented in terms of some optimal vector $\mathbf{r}$:

$$\mathbf{r} \approx \mathbf{a_1} - \mathbf{b_1} \approx \mathbf{a_2} - \mathbf{b_2} \quad (2)$$

The typical example used is

$$\mathbf{r} \approx king - man \approx queen - woman$$

and $\mathbf{r}$ approximately represents something like "is a royal version of". This can be rewritten as

$$king - man + woman \approx queen. \quad (3)$$

From this latter equation, the first standard word analogy test arises.

The prediction test and its dataset. In the prediction test, for the pairs of words in (1), evaluation proceeds by using the word analogy assumption

$$\mathbf{a_1} - \mathbf{b_1} + \mathbf{b_2} \approx \mathbf{a_2} \quad (4)$$

by means of showing that the left side of this equation (consisting of **premise vectors**) predicts–that is, it is closer to–the word represented by $\mathbf{a_2}$ (the **gold vector**) than to any other word in the vocabulary, according to some distance metric, which is generally accepted to be cosine similarity. The micro-averaged accuracy is then reported.

The test data for the prediction test consists of the MSR and GOOGLE datasets. The MSR dataset has 8000 analogy questions of morpho-syntactic nature and concerning adjectives, nouns and verbs.[1] The GOOGLE dataset consists of 19,544 analogy questions, across 14 relation types, half of which are semantic relations and half morpho-syntactic.

The ranking test and its dataset. In the ranking test, a list of word pairs is given that hold the same relation, but to differing degrees. The task is to rank these pairs by order of strength of the relation. Using the prediction test, this task requires the system to calculate the prediction for each pair of words $(\mathbf{a}, \mathbf{b})$ with respect the rest of the pairs on the list, and average these scores for $(\mathbf{a}, \mathbf{b})$. Pairs are ranked according to this average. The larger the average, the more typical a pair is predicted to be of the relation in question. Rankings are compared with a gold ranking by computing the Spearman's correlation rank coefficient.

The SEMEVAL 2012 Task 2 dataset is the standard word analogy ranking test test. It contains lists of pairs for 79 semantic relations.[2]

Implementation considerations. In our testing, out-of-vocabulary words were given the the component-wise average word embedding. It is important to note that in *all* test suites (also for those developed within embedding learning systems), we have found two conventional engineering choices: (1) normalisation of all word embeddings before testing, and (2) exclusion of the possibility to predict any premise vectors. We discuss these and other issues in the following section.

3 Problems with word analogy tests and empirical results

We identify three types of causes for concern when applying analogy testing, having to do with (1) a misalignment of assumptions in generating and testing word embeddings, (2) conventional engineering choices, and (3) problematic visual evidence derived from PCA for data projection to two dimensions.

For reasons of reproducibility, we downloaded and directly used all dimensionalities of GloVe pretrained word embeddings generated over a 2014 Wikpedia dump and the Gigaword corpus, combined for a 6 billion token corpus (Pennington et al., 2014).[3]. In the tests, embeddings for unknown words are replaced by the mean vectors. All tests are made using a version of a freely available embedding benchmarking software that we have extended for the purposes of this paper.[4]

[1] http://research.microsoft.com/en-us/um/people/gzweig/Pubs/myz_naacl13_test_set.tgz

[2] https://sites.google.com/site/semeval2012task2/

[3] Available at http://nlp.stanford.edu/projects/glove/

[4] https://github.com/natschluter/

3.1 Misalignment of assumptions

To date, methods for generating monolingual word embeddings purely from raw text, which make no use of hand-crafted or other lexical resources, nor any system of enrichment of the text, like parsers, POS-taggers or otherwise, have been based only on the distributional hypothesis: that words can be described sufficiently in terms of their distribution in language. Systems generating word embeddings in this manner use their generated representations to predict word contexts, or vice versa. So it is plausible that words that share much contextual information, and therefore much distributional information, will share similar representations and naturally group together in their hyperspace.

Supposing that such a word embedding generation system groups together words for humans from specific countries, like *Frenchman*, *Spaniard*, and *Dane*. We assume the same for the words *French*, *Spanish*, and *Danish*. While the system has probably successfully represented the distributional character of the words by grouping each set together, there is no reason why within each individual group, *Danish* and *Dane*'s relative positions should be similar to that of both pairs (*Spanish*, *Spain*) and (*French*, *Frenchman*). The assumption of distributional similarity does not align with the word analogy assumption.

In the extreme, we could theoretically have the pair (*Danish*, *Dane*)'s relative position most similar to that of the shuffled pairs (*French*, *Spaniard*) and (*Spanish*, *Frenchman*) and maintain identical word similarity scores on average. Indeed, one could shuffle the vector representations of all the words considered to be synonymous from the similarity benchmarking dataset; this would maintain precisely the same similarity score, using a cosine similarity metric.[5] Let $\pi : V \to V$ be a permutation of word vectors such that similar word remain close in the space. In particular, let's suppose that π shuffles the vectors of all nationalities, like *Dane* and *Frenchman*, but maintains the same language vectors like *Danish* and *French*. The average of similarities remains the same, as all terms appearing in the sum in (5) also appear in (6):

$$\frac{1}{n(n-1)/2} \sum_{\substack{i<j \\ i,j\in[n]}} \cos(\mathbf{a_i}, \mathbf{a_j}) \tag{5}$$

$$= \frac{1}{n(n-1)/2} \sum_{\substack{i<j \\ i,j\in[n]}} \cos(\pi(\mathbf{a_i}), \pi(\mathbf{a_j})) \tag{6}$$

However, the word analogy assumption is now most certainly broken: suppose that π permutes only two vectors, $\mathbf{a_2}$ and $\mathbf{a_3}$ and leaves all other vectors as is:

$$\mathbf{a_1} - \mathbf{b_1} + \mathbf{b_2} = \pi(\mathbf{a_2}) = \mathbf{a_3} \neq \mathbf{a_2}.$$

3.2 Conventional engineering choices

There are two conventional practices in evaluating word embeddings that we aim to show are problematic: normalisation and the exclusion of premise vectors in prediction.

Distortion by normalisation. It is common practice to normalise word embeddings before they are used, and in the case of word analogies, before they are tested. Unfortunately, this practice distorts the original spread of the word embeddings, which greatly effects testing for word analogies. In Table 1 we list the mean and variance of the norms of GloVe word vectors. We notice that on average the norm of the vectors is far from length 1, and the variance is so small that a large majority of vectors have length larger than 1. The word embedding learner was originally free to and would generally make use of a much larger portion of the hyperspace to discriminate based on word distribution.

d	mean	variance
50	4.475	0.744
100	3.977	0.847
300	4.966	1.471

Table 1: Spread of norms of GloVe word vectors across dimensions d.

We observe in Table 2 that scores change (and in fact drop) significantly when vectors are not normalised, for the GOOGLE and MSR tests. This suggests additionally that much of the success in analogy testing was misleading, resulting generally from collapsing the vocabulary of vectors onto

`word-analogy-caveat` extended from https://github.com/kudkudak/word-embeddings-benchmarks.

[5] This also works for a euclidean distance similarity metric.

the unit hypersphere. Any possible use of meaningful collinearity by the word-embedding model is lost after normalisation.

Exclusion of premise vectors from predictions. Another conventional practice in evaluating word embeddings by word analogy is the exclusion of premise vectors from the possibility of being predicted. As we can see in the results in Table 2, between 15-60% of the time, the system predicts a premise vector on the GOOGLE analogy data, for example. Upon closer analysis, we find that 99% of these latter prediction mis-hits are with the premise in the gold vector's own word pair; this means that words $\mathbf{a_1}, \mathbf{b_1}$ in word pairs are often so close together that they cancel each other out: $\mathbf{a_1} - \mathbf{b_1} \approx \mathbf{0}$. If the data truly scored high on the word analogy test, it would not need to exclude premise vectors from the possibility of prediction.

3.3 PCA to two dimensions from dimension d can be misleading

Results of the word analogy test are often accompanied by a visualisation of projected word vectors to the two dimensional plane using Principal Component Analysis (PCA) ((Mikolov et al., 2013a; Sun et al., 2015) for example). Though these are generally not claimed to be part of the evaluation, the visualisations are included to convince the reader of the quality of the word embeddings with respect to word analogies–the line connecting $\mathbf{a_1}$ and $\mathbf{b_1}$ being approximately parallel to the line through $\mathbf{a_2}$ and $\mathbf{b_2}$ whenever word analogy recovery is optimal (as in Equation (2)).

PCA is an unsupervised approach for finding the "core" features from the data, supposing a normal distribution feature-wise. For two dimensions, the objective is to find the two directions $\mathbf{e_1}, \mathbf{e_2}$ along which the data has the highest variability, and model the instances $x_k, k \in [N]$ by the respective distances a_{k1}, a_{k2} between the point x_k and lines through the mean vector, $\mathbf{m} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x_k}$ in the respective directions $\mathbf{e_1}$ and $\mathbf{e_2}$.

There are two main problems with this sort of evidence. Firstly, even if word analogies as described by Equation (4) existed in the data, it would only be a matter of chance that applying PCA to the entire dataset would recover even slightly these parallel (analogous) word relations visually. That is, there is no reason to believe that the line through the words in a pair is not almost perpendicular to the surface they are mapped to.

Secondly, if one is tempted to apply PCA only to the set of vectors corresponding to the two word groups in question, it is rather straightforward to produce the desired visualisation, so long as the two groups are clustered together. PCA should derive a surface that cuts through these two groups. So unless there is absolutely no clustering of similarly behaving words, PCA will give the evidence of word analogies one desires.

d		GOOGLE	MSR	SEMEVAL
50		46.24	35.56	13.99
	H	30.43	20.36	13.99
	D	20.58	10.01	14.76
	H,D	17.96	6.9	14.76
100		63.19	55.09	16.53
	H	33.47	24.87	16.53
	D	49.92	35.58	17.12
	H,D	34.44	18.06	17.12
300		71.85	61.64	17.0
	H	19.42	11.85	17.0
	D	65.32	51.58	16.91
	H,D	25.94	12.84	16.91

Table 2: Results of the word analogy tests, also without distortion through normalisation (D), without removing premise vectors from the set of possible gold vectors (H), and without either (H,D).

4 Concluding remarks

We have shown that there are serious problems with the appropriateness and informativeness of word analogy tests in current distributional word embedding evaluation. The first problem that should be addressed is the appropriateness. If word analogies are considered important enough, then word embedding generation systems should start to reflect this assumption. Until then, word analogies, as they are defined here, happen by rather chance. Once this assumption is built into systems, we still should put into question various details of the tests. Is a one-hit accuracy sufficiently informing on success in word analogy, or do we need a softer measure from for example the ranking world? These questions remain open for future work.

References

Zellig Harris. 1954. Distributional structure. *Word* 10:146–162.

David Jurgens, Saif Mohammad, Peter D. Turney, and Keith J. Holyoak. 2012. Semeval-2012 task 2: Measuring degrees of relational similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2012*. Montréal, Canada, pages 356–364.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proc of Coling*. pages 171–180.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

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

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar.

Fei Sun, Jiafeng Guo, Yanyan Lan, Jun Xu, and Xueqi Cheng. 2015. Learning word representations by jointly modeling syntagmatic and paradigmatic relations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Beijing, China, pages 136–145.