Evaluating word embeddings with a multi-categorical analogy task

Greg Bruss
Dept. of Data Science and Knowledge Engineering
Maastricht University
bruss.gregory@gmail.com

Abstract

Word embeddings and distributed representations present a powerful way of completing many NLP tasks, and capture fine-grained semantic and syntactic regularities. In this work, I evaluate the distributional space of these embeddings using a well-known analogy task over multiple categories with comparison to a human baseline, and find that performance is highly dependent on the category. An introduction to Global Vectors is also presented, and the analogy task is presented along with a complete implementation.

1 Credits

Distributed representations in the form of Word2Vec (Mikolov, 2013) have revolutionized the way the NLP community does many downstream tasks. In this paper, I follow the work of Turney (2013) and Gladkova (2016) in evaluating this distributed representation into a semantic space using a word analogy task. I make extensive use of pre-trained word embeddings using the GloVe architecture (Pennington, 2014) and comment on issues of evaluation highlighted by Linzen (2015)

2 Introduction

Distributed representations largely solved the problem of needing to use one-hot vectors to represent words. One-hot vectors, although easy to construct, are usually not helpful for many downstream NLP tasks. One of the major reasons is that one-hot vectors cannot accurately express the similarity between different words or groups of words, and lead to extremely sparse vectors when dealing with the large corpora that are helpful in NLP. Often, we want to make use of cosine similarities:

$$\frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} \in [-1, 1]$$

Since the cosine similarity of any two one-hot word vectors is 0, it is impossible to compare words in this way. Word2Vec and follow up distributed representations solve this issue by representing each word with a fixed length vector and embedding these vectors in a high dimensional space. Naturally, there is strong interest in evaluating these vector space models without constructing an entire NLP system around them (Linzen, 2015).

One such evaluation approach is the task of finding analogies. An analogy, broadly defined, is a linguistic relation along some axis. For example between the base form of a verb and its past tense (have and had), or a relation between a country of the world and its capital (Beijing and China). A method known as linear offset is well known to be able to encode these linguistic relations. One can estimate the offset using a certain relation as a starting point and then taking that relation and extending it to any new pair of words that are related in the same way. For the past tense example, [have, had] is linguistically related in the same way to [run, ran] in that they both differ only in the tense.

3 K-Nearest Neighbours

Word embeddings are continuous spaces. Thus any vector offset is unlikely to hit a word in the space exactly, thus we settle for taking the word closest to the point that we do hit.

Following the formal definition of Linzen (2014), we define the analogy task as being given by:

a:a* :: b :b* Where a and a* are the initial pair that defines the semantic relationship, b is the given word and b* is given by:

$$x^* = x' argmax \cos\left(x', a^* - a + b\right)$$

Where:

$$\cos(v, w) = \frac{v \cdot w}{\|v\|\|w\|}$$

4 Improving Word2Vec through Global Vectors

This work uses GloVe (Pennington, 2014) in the implementation. GloVe, short for Global Vectors, differs from other distributed word representations, mostly because of its ability to capture global corpus statistics. The following differences between GloVe and the skip-gram model of word2vec are relevant:
1. Two scalar model parameters are added for each word: the bias terms *bi* for (central target words) and *ci* (for context words)
2. The weight of each loss is replaced with a monotonically increasing function, namely $h\left(x_{ij}\right)$, whose range lies between [0,1]

5 Categorical Analogy Dataset

For this task, different categories of analogies were used, and were sourced from the Bigger Analogy Test Set (Gladkova, 2016). The reason for this is that multiple different categories encode different types of semantic relations, and measuring performance on these cross-representational tasks should give insight into the semantic space.The categories were chosen specifically to encode different types of relations. Furthermore, the Bigger Analogy Test Set has been designed to be balanced and representative, in that it covers both derivational morphology and encyclopaedic semantics - analogies that require knowledge about the world. Care has also been taken to reduce homonyms in the dataset - words which have the same sound but different meaning, such as the word "bat" which can be both a winged animal, and a wooden club used in baseball.
The analogy categories chosen for this paper were:

Country/Capital: A standard analogy such as London is to England as Paris is to France"

Present/Past Verb: A present test verb form along with its past tense form, such as Do is to did, as go is to went

Adjective/Superlative: An adjective followed by its superlative form, such as Big is to Biggest as Strong is to Strongest.

6 Model Architecture

A pre-trained, 50 dimensional GloVe word embedding was used for this task, which was trained on 6 billion tokens from the Wikipedia 2014 dataset. In extensions of this work, one could use different sized embeddings and see how performance differs. A 300-dimensional embedding trained on 42 billion tokens is available. Some tests were run with it, but the computational overhead was too large for this project.

7 Issues with Analogy Finding

The overall centrality of cosine similarity in this task is potentially a point of concern, in that it may be the case that the method is evaluating not just linguistic regularity but actually the local neighborhood of the x vector:

$$x = a^* - a + b$$

.
Following the characterization of Linzen: If for example a^* and a are very similar to each other, then the method might just return the nearest neighbour of b. This can be described as a form of *distance bias* in the distributional space. For any results, it will be important to note the evidence for this bias and to what extent it affects model performance.

8 Experimental Setup

25 examples were used for each of the three categories, sourced from the Bigger Analogy Test Set (Gladkova, 2015). These were evaluated using a 50 dimensional GloVe model described in section 3 and the method outlined in section 1. The task was also given to 5 people and their results are also included for a more valid comparison to be made.

Note that this is not supervised learning, and thus no analogies were fed to the model in the beginning. This is simply a one-shot way of evaluating an already trained embedding. Thus any correct answers are very unlikely to occur by chance. The GloVe model being used has 400 000 unique tokens, thus random guessing would almost certainly produce no correct answers.

9 The Distributional Space of the Word Embedding

Word embeddings naturally come in many different forms - the extent to which they differ depends

Category	Accuracy
country/capital	0.82
adjective-superlative	0.28
present tense -past tense	0.28

Table 1: GloVe model performance across the multi-categorical analogy task

Category	Accuracy
country/capital	0.95
adjective-superlative	1.0
present tense -past tense	1.0

Table 2: Human performance on the analogy task

on the training corpus used to populate the semantic space and the methods used to train them, for example the size of the context window (Mikolov, 2013) and whether or not downsampling is used.

10 Results

The results are given in tables 1 and 2. As can be seen, the model is in fact able to solve the analogy task in some sense, with strong performance on the capital/country relation, and worse performance on tasks of derivational morphology, such as present/past. It is worth noting that no machine system has ever achieved greater than 50% over the whole BATS dataset.

10.1 Discussion of Results

It is clear that human beings are better at the analogy task than this model. There are many reasons for this, some more likely than others. Analogy finding is a very broad task, and linguistic relationships have the potential to span large distances over the semantic space. Thus in order for our method to work, it needs to be extremely precise about where it lands in this high dimensional space. Humans appear to be very good at this. In addition, in many types of relationships, more than one answer is possible. Future work should investigate whether the top-5 accuracy markedly improves over the top-1 prediction, similar to how ImageNet (Deng, 2009) is evaluated.

10.2 Limitations

The following test is limited in that we only compared across 3 hand-picked categories. A different selection of 3 (or more) categories might give a different set of results. However, in this case, with the selection of 2 tasks of derivational morphology and a category of world knowledge (country/capital), the results do show where the model fails, which gives insight.

10.3 Conclusion

This work has presented a complete implementation of the analogy finding task over multiple categories for a pre-trained word embedding, and evaluated the results in comparison to a human baseline. It was shown that human beings completely outperform this model, but that the model does a fairly good job given the fact that random guessing would almost certainly produce no correct answers. Issues of general evaluation of semantic spaces was discussed, as well as potential markers of bias in the space - namely the distance bias introduced by the prominent role of cosine similarity.

Future work should look into quantitative evaluation of how the performance gap changes as we increase the size of the word vector, perhaps from 50-dimensional to 300-dimensional.

11 Supplemental Material

The data was preprocessed such that each line contains four terms. The first two terms, as in our analogy formulation, are a and $a*$, and the last two are b and $b*$. This allows the method described in section 3 to be used where we calculate x (the prediction of the fourth term) by subtracting the 2nd term from the first term and then adding the 3rd term.

12 References

Drozd, A. et al (2016) Word Embeddings, Analogies, and Machine Learning: Beyond King - Man + Woman = Queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Pages 35193530*

Gladkova, A. et al (2016) Analogy-based detection of morphological and semantic relations with word embeddings: what works andwhat doesn't. In *Proceedings of the NAACL Student Research Workshop, pages 8-15, San Diego, California, June. Association for Computational Linguistics*

Mikolov, T. et al (2013) Efficient estimation of word representations in vector space. In *Proceedings of ICLR*

Pennington, J. et al (2014), Glove: Global vectors for word representation. In *EMNLP, 14, page 1532–1543. (2014)*

Turney, P. et al (2010) From frequency to meaning: Vector space models of semantics. In *Journal of Artificial Intelligence Research, 2010*

Turney, P. et al (2013) Distributional Semantics BeyondWords: Supervised Learning of Analogy and Paraphrase. In *Transactions of the Association for Computational Linguistics, Volume 1, Pages 353366*, 2013.