Community Evaluation and Exchange of Word Vectors at wordvectors.org

Manaal Faruqui and **Chris Dyer**
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
{mfaruqui, cdyer}@cs.cmu.edu

Abstract

Vector space word representations are useful for many natural language processing applications. The diversity of techniques for computing vector representations and the large number of evaluation benchmarks makes reliable comparison a tedious task both for researchers developing new vector space models and for those wishing to use them. We present a website and suite of offline tools that that facilitate evaluation of word vectors on standard lexical semantics benchmarks and permit exchange and archival by users who wish to find good vectors for their applications. The system is accessible at: www.wordvectors.org.

1 Introduction

Data-driven learning of vector-space word embeddings that capture lexico-semantic properties is a technique of central importance in natural language processing. Using co-occurrence statistics from a large corpus of text (Deerwester et al., 1990; Turney and Pantel, 2010), it is possible to construct high-quality semantic vectors — as judged by both correlations with human judgements of semantic relatedness (Turney, 2006; Agirre et al., 2009) and as features for downstream applications (Turian et al., 2010). A number of approaches that use the internal representations from models of word sequences (Collobert and Weston, 2008) or continuous bags-of-context wordsets (Mikolov et al., 2013) to arrive at vector representations have also been shown to likewise capture co-occurrence tendencies and meanings.

With an overwhelming number of techniques to obtain word vector representations the task of comparison and choosing the vectors best suitable for a particular task becomes difficult. This is further aggravated by the large number of existing lexical semantics evaluation benchmarks being constructed by the research community. For example, to the best of our knowledge, for evaluating word similarity between a given pair of words, there are currently at least 10 existing benchmarks[1] that are being used by researchers to prove the effectiveness of their word vectors.

In this paper we describe an online application that provides the following utilities:

- Access to a suite of word similarity evaluation benchmarks
- Evaluation of user computed word vectors
- Visualizing word vectors in $\mathbb{R}^2$
- Evaluation and comparison of the available open-source vectors on the suite
- Submission of user vectors for exhaustive offline evaluation and leader board ranking
- Publicly available repository of word vectors with performance details

Availability of such an evaluation system will help in enabling better consistency and uniformity in evaluation of word vector representations as well as provide an easy to use interface for end-users in a similar spirit to Socher et al. (2013a), a website for text classification.[2] Apart from the online demo version, we also provide a software that can be run in an offline mode on the command line. Both the online and offline tools will be kept updated with continuous addition of new relevant tasks and vectors.

[1] www.wordvectors.org/suite.php
[2] www.etcml.com

2 Word Similarity Benchmarks

We evaluate our word representations on 10 different benchmarks that have been widely used to measure word similarity. The first one is the **WS-353**[3] dataset (Finkelstein et al., 2001) containing 353 pairs of English words that have been assigned similarity ratings by humans. This data was further divided into two fragments by Agirre et al. (2009) who claimed that *similarity* (**WS-SIM**) and *relatedness* (**WS-REL**)[4] are two different kinds of relations and should be dealt with separately. The fourth and fifth benchmarks are the **RG-65** (Rubenstein and Goodenough, 1965) and the **MC-30** (Miller and Charles, 1991) datasets that contain 65 and 30 pairs of nouns respectively and have been given similarity rankings by humans. These differ from **WS-353** in that it contains only nouns whereas the former contains all kinds of words.

The sixth benchmark is the **MTurk-287**[5] (Radinsky et al., 2011) dataset that constitutes 287 pairs of words and is different from the previous benchmarks in that it has been constructed by crowdsourcing the human similarity ratings using Amazon Mechanical Turk (AMT). Similar in spirit is the **MTruk-771**[6] (Halawi et al., 2012) dataset that contains 771 word pairs whose similarity was crowdsourced from AMT. Another, AMT created dataset is the **MEN**[7] benchmark (Bruni et al., 2012) that consists of 3000 word pairs, randomly selected from words that occur at least 700 times in the freely available ukWaC and Wackypedia[8] corpora combined.

The next two benchmarks were created to put emphasis on different kinds of word types. To specifically emphasize on verbs, Yang and Powers (2006) created a new benchmark **YP-130** of 130 verb pairs with human similarity judgements. Since, most of the earlier discussed datasets contain word pairs that are relatively more frequent in a corpus, Luong et al. (2013) create a new benchmark (**Rare-Word**)[9] that contains rare-words by sampling words from different frequency bins to a total of 2034 word pairs.

We calculate similarity between a given pair of words by the *cosine* similarity between their corresponding vector representation. We then report Spearman's rank correlation coefficient (Myers and Well, 1995) between the rankings produced by our model against the human rankings.

Multilingual Benchmarks. As is the case with most NLP problems, the lexical semantics evaluation benchmarks for languages other than English have been limited. Currently, we provide a link to some of these evaluation benchmarks from our website and in future will expand the website to encompass vector evaluation for other languages.

3 Visualization

The existing benchmarks provide ways of vector evaluation in a quantitative setting. To get an idea of what kind of information the vectors encode it is important to see how these vectors represent words in n-dimensional space, where n is the length of the vector. Visualization of high-dimensional data is an important problem in many different domains, and deals with data of widely varying dimensionality. Over the last few decades, a variety of techniques for the visualization of such high-dimensional data have been proposed (de Oliveira and Levkowitz, 2003).

Since visualization in n dimensions is hard when $n >= 3$, we use the t-SNE (van der Maaten and Hinton, 2008) tool[10] to project our vectors into $\mathbb{R}^2$. t-SNE converts high dimensional data set into a matrix of pairwise similarities between individual elements and then provides a way to visualize these distances in a way which is capable of capturing much of the local structure of the high-dimensional data very well, while also revealing global structure such as the presence of clusters at several scales.

In the demo system, we give the user an option to input words that they need to visualize which are fed to the t-SNE tool and the produced images are shown to the user on the webpage. These images can then be downloaded and used. We have

[3] http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/

[4] http://alfonseca.org/eng/research/wordsim353.html

[5] http://tx.technion.ac.il/~kirar/Datasets.html

[6] http://www2.mta.ac.il/~gideon/mturk771.html

[7] http://clic.cimec.unitn.it/~elia.bruni/MEN.html

[8] http://wacky.sslmit.unibo.it/doku.php?id=corpora

[9] http://www-nlp.stanford.edu/~lmthang/morphoNLM/

[10] http://homepage.tudelft.nl/19j49/t-SNE_files/tsne_python.zip

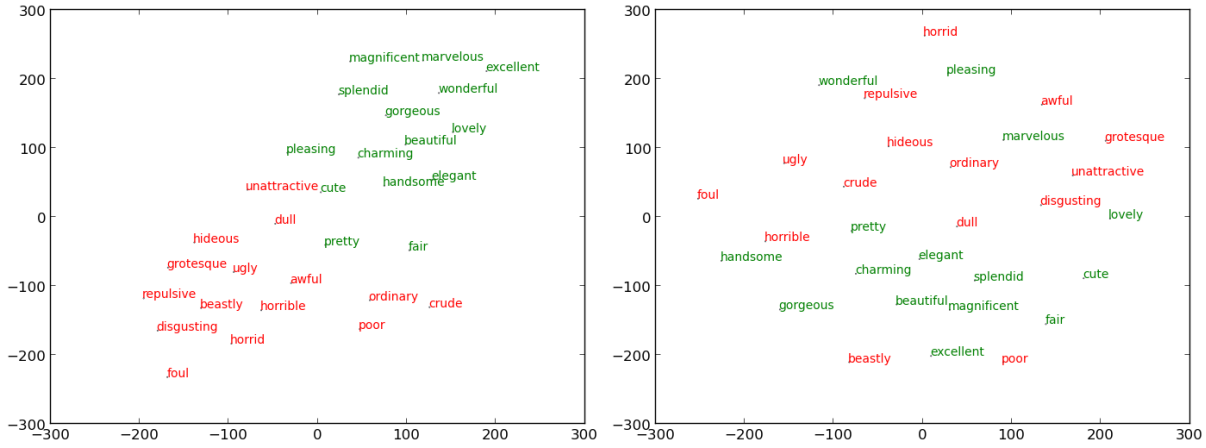

Figure 1: Antonyms (red) and synonyms (green) of *beautiful* represented by Faruqui and Dyer (2014) (left) and Huang et al. (2012) (right).

included two datasets by default which exhibit different properties of the language:

- Antonyms and synonyms of *beautiful*
- Common male-female nouns and pronouns

In the first plot, ideally the antonyms (*ugly, hideous,* ...) and synonyms (*pretty, gorgeous,* ...) of beautiful should form two separate clusters in the plot. Figure 1 shows the plots of the antonyms and synonyms of the word *beautiful* for two available embeddings. The second default word plot is the gender data set, every word in which has a male and a female counterpart (ex. *grandmother* and *grandfather*), this data set exhibits both local and global properties. Locally, the male and female counterparts should occur in pairs together and globally there should be two separate clusters of male and female.

4 Word Vector Representations

4.1 Pre-trained Vectors

We haves collected several standard pre-trained word vector representations freely available for research purposes and provide a utility for the user to test them on the suite of benchmarks, as well as try out the visualization functionality. The user can also choose the option to choose two different types of word vectors and compare their performance on the benchmarks. We will keep adding word vectors on the website as and when they are released. The following word vectors have been included in our collection:

Metaoptimize. These word embeddings [11] have been trained in (Turian et al., 2010) using a neural network language model and were shown to be useful for named entity recognition (NER) and phrase chunking.

SENNA. It is a software[12] which outputs a host of predictions: part-of-speech (POS) tags, chunking, NER etc (Collobert et al., 2011). The software uses neural word embeddings trained over Wikipedia data for over 2 months.

RNNLM. The recurrent neural network language modeling toolkit[13] comes with some pre-trained embeddings on broadcast news data (Mikolov et al., 2011).

Global Context. Huang et al. (2012) present a model to incorporate document level information into embeddings to generate semantically more informed word vector representations. These embeddings[14] capture both local and global context of the words.

Skip-Gram. This model is a neural network language model except for that it does not have a hidden layer and instead of predicting the target word, it predicts the context given the target word (Mikolov et al., 2013). These embeddings are much faster to train[15] than the other neural embeddings.

[11] http://metaoptimize.com/projects/wordreprs/
[12] http://ronan.collobert.com/senna/
[13] http://rnnlm.org/
[14] http://nlp.stanford.edu/~socherr/ACL2012_wordVectorsTextFile.zip
[15] https://code.google.com/p/word2vec/

Select?	Name	Dimensions	Vocabulary	Reference
☐	Metaoptimize	50	268810	Turian et al, 2010
☐	Senna	50	130000	Collobert et al, 2011
☐	RNN	80	82390	Mikolov et al, 2011
☐	RNN	640	82390	Mikolov et al, 2011
☐	Global Context	50	100232	Socher et al, 2012
☐	Skip-Gram	640	180834	Mikolov et al 2013
☐	Multilingual	512	180834	Faruqui and Dyer, 2014

Figure 2: Vector selection interface (right) of the demo system.

Multilingual. Faruqui and Dyer (2014) propose a method based on canonical correlation analysis to produce more informed monolingual vectors using multilingual knowledge. Their method is shown to perform well for both neural embeddings and LSA (Deerwester et al., 1990) based vectors.[16]

4.2 User-created Vectors

Our demo system provides the user an option to upload their word vectors to perform evaluation and visualization. However, since the size of the word vector file will be huge due to a lot of infrequent words that are not useful for evaluation, we give an option to filter the word vectors file to only include the words required for evaluation. The script and the vocabulary file can be found on the website online.

5 Offline Evaluation & Public Access

We provide an online portal where researchers can upload their vectors which are then be evaluated on a variety of NLP tasks and then placed on the leader board.[17] The motivation behind creating such a portal is to make it easier for a user to select the kind of vector representation that is most suitable for their task. In this scenario, instead of asking the uploader to filter their word vectors for a small vocabulary, they will be requested to upload their vectors for the entire vocabulary.

[16] http://cs.cmu.edu/~mfaruqui/soft.html

[17] We provide an initial list of some such tasks to which we will later add more tasks as they are developed.

5.1 Offline Evaluation

Syntactic & semantic relations. Mikolov et al. (2013) present a new semantic and syntactic relation dataset composed of analogous word pairs of size 8869 and 10675 pairs resp.. It contains pairs of tuples of word relations that follow a common relation. For example, in *England : London :: France : Paris*, the two given pairs of words follow the country-capital relation. We use the vector offset method (Mikolov et al., 2013) to compute the missing word in these relations. This is non-trivial $|V|$-way classification task where V is the size of the vocabulary.

Sentence Completion. The Microsoft Research sentence completion dataset contains 1040 sentences from each of which one word has been removed. The task is to correctly predict the missing word from a given list of 5 other words per sentence. We average the word vectors of a given sentence $q_{sent} = \sum_{i=1,i\neq j}^{N} q_{w_i}/N$, where w_j is the missing word and compute the cosine similarity of q_{sent} vector with each of the options. The word with the highest similarity is chosen as the missing word placeholder.

Sentiment Analysis Socher et al. (2013b) have created a treebank which contains sentences annotated with fine-grained sentiment labels on both the phrase and sentence level. They show that compositional vector space models can be used to predict sentiment at these levels with high accuracy. The coarse-grained treebank, containing only positive and negative classes has been split into training, development and test datasets con-

Serial	Dataset	Num Pairs	Not found	Rho
1	EN-MC-30.txt	30	0	0.8198
2	EN-MTurk-287.txt	287	1	0.5365
3	EN-RG-65.txt	65	0	0.7554
4	EN-RW-STANFORD.txt	2034	598	0.4232
5	EN-WS-353-ALL.txt	353	0	0.6809
6	EN-WS-353-REL.txt	252	0	0.6462
7	EN-WS-353-SIM.txt	203	0	0.7440
8	EN-MEN-TR-3k.txt	3000	1	0.7585

Figure 3: Screenshot of the command line version showing word similarity evaluation.

taining 6920, 872 and 1821 sentences respectively. We train a logistic regression classifier with $L2$ regularization on the average of the word vectors of a given sentence to predict the coarse-grained sentiment tag at the sentence level.

TOEFL Synonyms. These are a set of 80 questions compiled by Landauer and Dutnais (1997), where a given word needs to be matched to its closest synonym from 4 given options. A number of systems have reported their results on this dataset.[18] We use cosine similarity to identify the closest synonym.

5.2 Offline Software

Along with the web demo system we are making available a software which can be downloaded and be used for evaluation of vector representations offline on all the benchmarks listed above. Since, we cannot distribute the evaluation benchmarks along with the software because of licensing issues, we would give links to the resources which should be downloaded prior to using the software. This software can be run on a command line interface. Figure 3 shows a screenshot of word similarity evaluation using the software.

5.3 Public Access

Usually corpora that the vectors are trained upon are not available freely because of licensing issues but it is easier to release the vectors that have been trained on them. In the system that we have developed, we give the user an option to either make the vectors freely available for everyone to use under a GNU General Public License[19] or a Creative Commons License.[20] If the user chooses not to make the word vectors available, we would evaluate the vectors and give it a position in the leader board with proper citation to the publications/softwares.

[18] http://aclweb.org/aclwiki/index.php?title=TOEFL_Synonym_Questions_(State_of_the_art)

[19] https://www.gnu.org/copyleft/gpl.html

[20] https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode

6 Conclusion

In this paper we have presented a demo system that supports rapid and consistent evaluation of word vector representations on a variety of tasks, visualization with an easy-to-use web interface and exchange and comparison of different word vector representations. The system also provides access to a suite of evaluation benchmarks both for English and other languages. The functionalities of the system are aimed at: (1) Being a portal for systematic evaluation of lexical semantics tasks that heavily rely on word vector representation, (2) Making it easier for an end-user to choose the most suitable vector representation schema.

Acknowledgements

We thank members of Noah's Ark and c-lab for their helpful comments about the demo system. Thanks to Devashish Thakur for his help in setting up the website. This work was supported by the NSF through grant IIS-1352440.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of NAACL*, NAACL '09, pages 19–27, Stroudsburg, PA, USA.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 136–145, Jeju Island, Korea, July. Association for Computational Linguistics.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep

neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 160–167, New York, NY, USA. ACM.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537, November.

Maria Cristina Ferreira de Oliveira and Haim Levkowitz. 2003. From visual data exploration to visual data mining: A survey. *IEEE Trans. Vis. Comput. Graph.*, 9(3):378–394.

S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, April. Association for Computational Linguistics.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: the concept revisited. In *WWW '01: Proceedings of the 10th international conference on World Wide Web*, pages 406–414, New York, NY, USA. ACM Press.

Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. 2012. Large-scale learning of word relatedness with constraints. In *KDD*, pages 1406–1414.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th ACL: Long Papers-Volume 1*, pages 873–882.

Thomas K Landauer and Susan T. Dutnais. 1997. A solution to platos problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, pages 211–240.

Minh-Thang Luong, Richard Socher, and Christopher D. Manning. 2013. Better word representations with recursive neural networks for morphology. In *CoNLL*, Sofia, Bulgaria.

Tomas Mikolov, Stefan Kombrink, Anoop Deoras, Lukar Burget, and J Cernockỳ. 2011. Rnnlm–recurrent neural network language modeling toolkit. *Proc. of the 2011 ASRU Workshop*, pages 196–201.

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

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Jerome L. Myers and Arnold D. Well. 1995. *Research Design & Statistical Analysis*. Routledge, 1 edition, June.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 337–346, New York, NY, USA. ACM.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Commun. ACM*, 8(10):627–633, October.

Richard Socher, Romain Paulus, Bryan McCann, Kai Sheng Tai, and Andrew Y. Hu, JiaJi Ng. 2013a. etcml.com - easy text classification with machine learning. In *Advances in Neural Information Processing Systems (NIPS 2013)*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013b. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Stroudsburg, PA, October. Association for Computational Linguistics.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th ACL*, ACL '10, pages 384–394, Stroudsburg, PA, USA.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning : Vector space models of semantics. *Journal of Artificial Intelligence Research*, pages 141–188.

Peter D. Turney. 2006. Similarity of semantic relations. *Comput. Linguist.*, 32(3):379–416, September.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, November.

Dongqiang Yang and David M. W. Powers. 2006. Verb similarity on the taxonomy of wordnet. In *In the 3rd International WordNet Conference (GWC-06), Jeju Island, Korea*.