A critique of word similarity as a method for evaluating distributional semantic models

Miroslav Batchkarov, Thomas Kober, Jeremy Reffin, Julie Weeds and **David Weir**
Text Analysis Group, Department of Informatics, University of Sussex
{M.Batchkarov, T.Kober, J.P.Reffin, J.E.Weeds, D.J.Weir}@sussex.ac.uk

Abstract

This paper aims to re-think the role of the word similarity task in distributional semantics research. We argue while it is a valuable tool, it should be used with care because it provides only an approximate measure of the quality of a distributional model. Word similarity evaluations assume there exists a single notion of similarity that is independent of a particular application. Further, the small size and low inter-annotator agreement of existing data sets makes it challenging to find significant differences between models.

1 Introduction

Distributional models of lexical semantics have recently attracted considerable interest in the NLP community. With the increase in popularity, the issue of evaluation is becoming more important. While extrinsic (task-based) evaluations are increasingly common, the most frequently used family of evaluation procedures (intrinsic evaluations) attempt to directly measure the "inherent" quality of a word representation. This often takes the form of computing the extent to which a model agrees with human-provided word or phrase similarity scores.

This paper highlights the theoretical and practical issues with the word similarity task, which make it a poor measure of the quality of a distributional model. We investigate five commonly used word similarity datasets, `RG` (Rubenstein and Goodenough, 1965), `MC` (Miller and Charles, 1991), `WS353` (Finkelstein et al., 2001), `MEN` (Bruni et al., 2014) and `SimLex` (Hill et al., 2015). Our contributions are as follows. We argue that the notion of lexical similarity is difficult to define outside of the context of a task and without conflating different concepts such as "similarity" or "relatedness". We show inter-annotator agreement at the word similarity task is considerably lower compared to other tasks such as document classification or textual entailment. Furthermore, we demonstrate that the quality of a model, as measured by a given word similarity data set, can vary substantially because of the small size of the data set. Lastly, we introduce a simple sanity check for word similarity data sets that tests whether a data set is able to reliably identify corrupted word vectors. These findings can be adopted as guidelines for designers of evaluation data sets. The code for our experiments is available at `github.com/mbatchkarov/repeval2016`.

2 Definition of Similarity

The notion of similarity is challenging to define precisely. Existing word similarity data sets typically contain a broad range of semantic relations such as synonymy, antonymy, hypernymy, co-hypernymy, meronymy and topical relatedness. Earlier word similarity work such as `WS353` does not attempt to differentiate between those. In contrast, more recent work such as `MEN` and `SimLex` distinguishes between "similarity" and "relatedness" and provide human annotators with more specific instructions as to what makes words similar.

However, all data sets considered in this paper assume that there exists a single gold-standard score for each pair of words, which can vary considerably across data sets, depending on what notion of similarity is used. For example, the pair "chicken–rice" has a normalised score of 0.14 in `SimLex` and 0.68 in `MEN`, while "man–woman" scores 0.33 and 0.84 respectively.

We argue that every downstream application de-

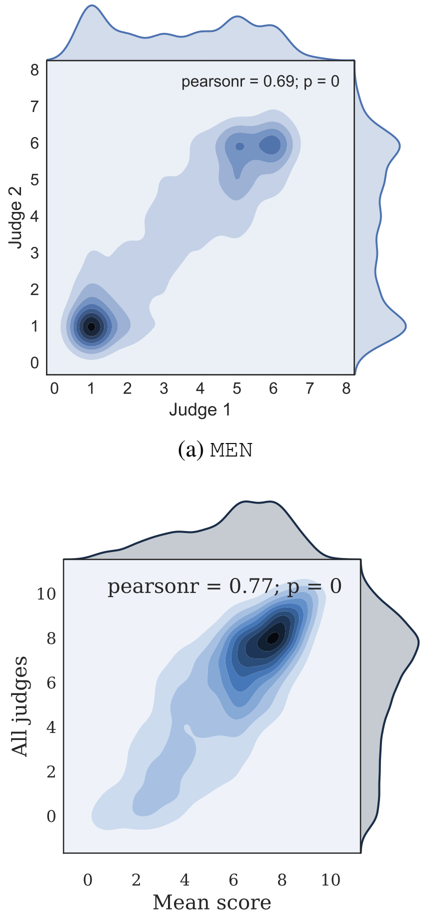
Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP, pages 7–12,
Berlin, Germany, August 12, 2016. ©2016 Association for Computational Linguistics

fines its own kind of similarity. Words are therefore not inherently similar or dissimilar. For example, "good acting" and "cluttered set" are highly dissimilar in terms of the sentiment they express towards a theatrical play. However, they are very similar in the context of detecting news items related to the theatre, as both phrases are highly indicative of theatre-related content. It is often unclear what kind of similarity is useful for a downstream problem in advance. Indeed, it has been shown that being able to learn the notion defined by a particular word similarity task does not necessarily translate to superior extrinsic performance (Schnabel et al., 2015). This argument parallels that of von Luxburg et al. (2012, p 2), who argue that "[d]epending on the use to which a clustering is to be put, the same clustering can either be helpful or useless". The quality of an unsupervised algorithm can therefore only be assessed in the context of an application.

3 Subjectivity and task difficultly

When human judges annotate word pairs for similarity, the distinctions in meaning they are asked to make are often very subtle, especially in the absence of context. For instance, the normalised similarity scores provided by 13 annotators for the pair "tiger–cat" range from 0.5 to 0.9 in WS353. This results in low inter-annotator agreement even between native speakers. This section analyses the variation in similarity scores produced by different annotators, and compares the agreement score for the first 13 annotators of WS353 and the two authors of MEN to typical agreements reported in the NLP literature for tasks such as document classification and textual entailment.

Figure 1 shows a kernel density estimate of the distribution of similarity scores between judges for MEN and WS353[1]. Both data sets exhibit undesirable characteristics. The distribution of scores assigned by both judges in MEN appears to be bimodal, which suggests that the annotators are operating on a three-point scale rather than on a seven-point one. There is also a significant amount of variation—the similarity assigned to a word pair exceeds two points (out of seven) in 313 cases[2] (10.4%) and can vary by as many as six points. WS353 exhibits a strong bias towards high-similarity word pairs. However, individual judges exhibit a bias towards similarity scores in the middle of the spectrum. Variance is also high — 535 individual annotations (10.3% of all cases) for a given word pair differ by more than two points (out of ten) from the mean score for that pair[3].



(a) MEN

(b) WS353

Figure 1: Distribution of similarity scores between annotators

It is not possible to compare inter-annotator agreement scores for word similarity and other natural-language labelling tasks directly. Labels in the former are on an ordinal scale, so agreement is measured using Spearman's rho (ρ). In contrast, the labels in other tasks are often categorical; agreement is typically measured using Cohen's kappa (κ). To address this issue, we convert word similarity scores to discrete labels by placing the continuous scores into equally sized bins. For example, the range of similarity scores in WS353

[1] The other data sets used in this study do not provide the annotations of each individual subject.

[2] MEN contains a total of 3000 annotated pairs.

[3] WS353 contains 353 unique word pairs annotated by at least 13 judges for a total of 5189 (word pair, annotation) units.

is $[0, 10]$, and the bin boundaries are at $[0, 5, 10]$ when using two bins and at $[0, 3.33, 6.66, 10]$ when using three bins. The three-item continuous labelling $[2.1, 5.8, 7.9]$ is converted to $[A, B, B]$ when using two bins and to $[A, B, C]$ when using three bins.

This conversion process suffers from two drawbacks. First, order information is lost, so misplacing an item in bin A instead of in bin B is considered as severe an error as misplacing an item from bin A into bin F. This is less of an issue when the bin count is small. Second, the number of bins is a free parameter ranging between 1 (all items in the same bin) and 7 or 10 (all items in original bins)[4]. κ is a decreasing function of the number of bins because it becomes harder for annotators to agree when there is a large number of bins to choose from. This analysis is agnostic as to how many bins should be used. We experiment with values between 2 and 5.

The inter-annotator agreement of WS353 and MEN (converted to Cohen's κ) is shown in Figure 2. Because κ is only applicable when there are exactly two annotators, we report an average κ over all pairwise comparisons[5]. A κ score can be computed between each of the 91 pairs of judges ("WS353-P"), or between each judge and the mean across all judges ("WS353-M") (as in Hill et al. (2015, Section 2.3)). Mean agreement ranges from $\kappa = 0.21$ to $\kappa = 0.62$.

For comparison, Kim and Hovy (2004) report $\kappa = 0.91$ for a binary sentiment task. Gamon et al. (2005) report a κ of 0.7–0.8 for a three-way sentiment task. Wilson et al. (2005) report $\kappa = 0.72$ for a four-class short expressions sentiment task, rising to $\kappa = 0.84$ if phrases marked as "unsure" are removed. McCormick et al. (2008) report $\kappa = 0.84$ for a five-way text classification task. Stolcke et al. (2000) report $\kappa = 0.8$ for a 42-label dialogue act tagging task. Toledo et al. (2012) achieve $\kappa = 0.7$ for a textual entailment task, and Sammons et al. (2010) report $\kappa = 0.8$ to $\kappa = 1$ for a domain identification task. All these κ scores are considerably higher than those achieved by WS353 and MEN.

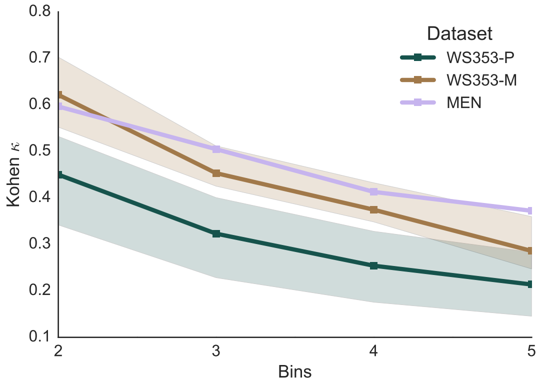


Figure 2: Inter-annotator agreement of WS353, measured in Cohen's κ. Shaded region shows the mean and one standard deviation around it. A standard deviation is not shown for MEN as only the annotation of a single pair of raters are available.

4 Size of data set

Another issue with existing word similarity data sets is their small size. This ranges from 30 to 3000 data points (Miller and Charles, 1991; Rubenstein and Goodenough, 1965; Landauer and Dumais, 1997; Finkelstein et al., 2001; Hill et al., 2015; Huang et al., 2012; Luong et al., 2013; Bruni et al., 2014). Moreover, they only feature a "tidy" subset of all naturally occurring words, free of spelling variation, domain-specific terminology and named entities. The focus is predominantly on relatively high-frequency words, so the quality of the model cannot be quantified fully. In contrast, typical distributional models "in the wild" have a vocabulary of tens or hundreds of thousands of types.

For practical applications, users need to understand the entire distributional model, not just the small fraction of it covered by an intrinsic evaluation. A side effect of using small evaluation data sets is that the measured correlation scores may vary significantly. However, variance is seldom reported in the literature. To quantify it, we train a word2vec model (Mikolov et al., 2013) on a mid-2011 copy of English Wikipedia. We use the CBOW objective with negative sampling and a window size of 5, as implemented in gensim (Řehůřek and Sojka, 2010). The model is evaluated on five word similarity data sets—MC, RG, WS353, SimLex and MEN. We compute the empirical distribution of correlation with human scores by bootstrapping. Each data set is resampled 500 times with replacement. The distri-

[4] WS353 was annotated on a ten-point scale, whereas MEN used a seven-point scale.

[5] Averaging is only needed for WS353, which has been annotated by (at least) 13 judges. MEN only provides full annotations for two judges.

butional model is evaluated on each sample (Efron and Tibshirani, 1994). Results are shown in Table 1a. We also evaluate a baseline model that represents words as completely random vectors, sampled from continuous uniform distribution $\mathcal{U}(0, 1)$ (Table 1b).

Dataset	Mean	Std	Min	Max	Size
MC	0.71	0.12	0.29	0.95	30
RG	0.72	0.06	0.50	0.87	65
WS353	0.64	0.04	0.53	0.75	353
SimLex	0.31	0.03	0.23	0.39	999
MEN	0.67	0.01	0.64	0.70	3000

(a) word2vec vectors

Dataset	Mean	Std	Min	Max	Size
MC	-0.01	0.19	-0.53	0.55	30
RG	0.08	0.11	-0.28	0.41	65
WS353	-0.08	0.05	-0.24	0.10	353
SimLex	0.01	0.03	-0.09	0.12	999
MEN	-0.02	0.02	-0.08	0.04	3000

(b) Random vectors

Table 1: Distribution of Spearman ρ between model predictions and gold standard data set.

The mean correlation is in line with values reported in the literature. However, standard deviation is strongly dependent on the size of the gold-standard data set. Even for MEN, which is the largest word similarity data set in this study, the measured correlation varies as much as 0.06. However, this fact is not often addressed in the literature. For instance, the difference between the recently proposed Swivel (Shazeer et al., 2016) and word2vec with negative sampling is less than 0.02 on WS353, SimLex and MEN. Table 1 suggests that these differences may well not be statistically significant.

5 Sensitivity to noise

In this section we propose a simple sanity check for word similarity data sets, which we suggest is used periodically while developing a data set. It is based on the requirement that for a given evaluation method, good word representations should perform measurably better than poor ones. One method of reliably generating poor word vectors is to start with a distributional model and decrease its quality by adding random noise. The evaluation framework should be able to detect the difference between the original and corrupted models. Model performance, as measured by the evaluation method, should be a monotonically decreasing function of the amount of noise added. In the extreme case, a completely random distributional model should achieve a correlation of zero with the human-provided intrinsic similarity scores (Table 1b).

Figure 3 shows an application of our proposal to MC, RG and MEN. We add uniform random noise $\mathcal{U}(-n, n)$ to all elements of all word vectors from Section 4, where $n \in [0, 3]$. This is a considerable perturbation as the word vectors used have have a mean L2 norm of 2.4. RG and MC do not sufficiently capture the degradation of vector quality as noise is added because ρ may increase with n. The variance of the measurements is also very high. Both datasets therefore fail the sanity check. MEN's performance is considerably better, with smaller standard deviation and correlation tending to zero as noise is added. WS353 and SimLex exhibit similar behaviour to MEN, but have higher variance.

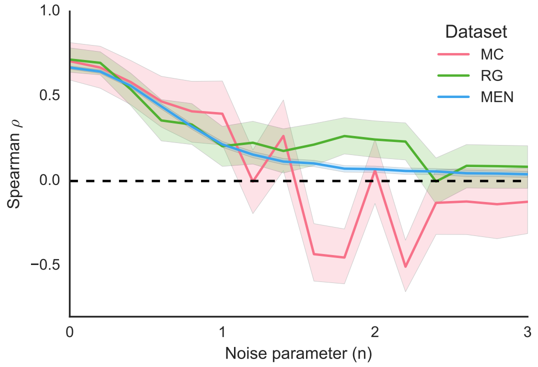


Figure 3: Word similarity noise test. Shaded region shows one standard deviation around the mean, which is estimated via bootstrapping.

6 Conclusion

This paper showed the word similarity task is considerably more challenging for annotators than extrinsic tasks such as document classification. Further, the small size of existing word similarity data sets results in high variance, making it difficult to reliably differentiate between models. More fundamentally, the task assumes there exists a single similarity score between a pair of words which is independent of a particular application. These results challenge the value of intrinsic data sets as a gold standard. We argue that word similarity still has a place in NLP, but researchers should be aware of its limitations. We view the task as a computationally efficient approximate measure of model quality, which can be useful in the early stage of model development. However, research should place less emphasis on word similarity performance and more on extrinsic results such as (Batchkarov, 2015; Huang and Yates, 2009; Milajevs et al., 2014; Schnabel et al., 2015; Turian et al., 2010; Weston et al., 2015).

Acknowledgements

This work was funded by UK EPSRC project EP/IO37458/1 *"A Unified Model of Compositional and Distributional Compositional Semantics: Theory and Applications"*. We would like to thank all members of the TAG lab team and our anonymous reviewers for their helpful comments.

References

Miroslav Batchkarov. 2015. *Evaluating distributional models of compositional semantics*. Ph.D. thesis, University of Sussex.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Proceedings of JAIR* 49:1–47.

Bradley Efron and Robert Tibshirani. 1994. *An Introduction to the Bootstrap*. CRC press.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. *Proceedings of the 10th international conference on World Wide Web* pages 406–414.

Michael Gamon, Anthony Aue, Simon Corston-Oliver, and Eric Ringger. 2005. Pulse: Mining customer opinions from free text. In *Advances in Intelligent Data Analysis VI*, Springer, pages 121–132.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics* .

Eric Huang, Richard Socher, Christopher Manning, and Andrew Ng. 2012. Improving word representations via global context and multiple word prototypes. *Proceedings of ACL* pages 873–882.

Fei Huang and Alexander Yates. 2009. Distributional representations for handling sparsity in supervised sequence-labeling. *Proceedings of ACL-IJCNLP* pages 495–503.

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. *Proceedings of ACL* page 1367.

Thomas Landauer and Susan Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104(2):211.

Minh-Thang Luong, Richard Socher, and Christopher Manning. 2013. Better word representations with recursive neural networks for morphology. *Proceedings of CoNLL* 104.

Patrick McCormick, Noémie Elhadad, and Peter Stetson. 2008. Use of semantic features to classify patient smoking status. *AMIA Annual Symposium Proceedings* 2008.

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

Dmitrijs Milajevs, Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, and Matthew Purver. 2014. Evaluating neural word representations in tensor-based compositional settings. *Proceedings of EMNLP* pages 708–719.

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

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* pages 45–50.

Herbert Rubenstein and John Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM* 8(10):627–633.

Mark Sammons, V. G. Vinod Vydiswaran, and Dan Roth. 2010. Ask not what textual entailment can do for you... In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 1199–1208.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings .

Noam Shazeer, Ryan Doherty, Colin Evans, and Chris Waterson. 2016. Swivel: Improving embeddings by noticing what's missing. *arXiv preprint arXiv:1602.02215* .

Andreas Stolcke, Noah Coccaro, Rebecca Bates, Paul Taylor, Carol Van Ess-Dykema, Klaus Ries, Elizabeth Shriberg, Daniel Jurafsky, Rachel Martin, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics* 26(3):339–373.

Assaf Toledo, Sophia Katrenko, Stavroula Alexandropoulou, Heidi Klockmann, Asher Stern, Ido Dagan, and Yoad Winter. 2012. Semantic annotation for textual entailment recognition. In *Advances in Computational Intelligence*, Springer, pages 12–25.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. *Proceedings of ACL* pages 384–394.

Ulrike von Luxburg, Robert Williamson, and Isabelle Guyon. 2012. Clustering: Science or art? *ICML Unsupervised and Transfer Learning* pages 65–80.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2015. Towards AI-complete question answering: A set of prerequisite toy tasks. *arXiv Preprint arXiv:1502.05698* .

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of HLT-EMNLP* pages 347–354.