Intrinsic Evaluations of Word Embeddings: What Can We Do Better?

Anna Gladkova
Department of Language
and Information Sciences
The University of Tokyo
Tokyo, Japan
gladkova@phiz.c.u-tokyo.ac.jp

Aleksandr Drozd
Global Scientific Information
and Computing Center
Tokyo Institute of Technology
Tokyo, Japan
alex@smg.is.titech.ac.jp

Abstract

This paper presents an analysis of existing methods for the intrinsic evaluation of word embeddings. We show that the main methodological premise of such evaluations is "interpretability" of word embeddings: a "good" embedding produces results that make sense in terms of traditional linguistic categories. This approach is not only of limited practical use, but also fails to do justice to the strengths of distributional meaning representations. We argue for a shift from abstract ratings of word embedding "quality" to exploration of their strengths and weaknesses.

1 Introduction

The number of word embeddings is growing every year. A new model is typically evaluated across several tasks, and is considered an improvement if it achieves better accuracy than its predecessors. There are numerous real-use applications that can be used for this purpose, including named entity recognition (Guo et al., 2014), semantic role labeling (Chen et al., 2014), and syntactic parsing (Chen and Manning, 2014).

However, different applications rely on different aspects of word embeddings, and good performance in one application does not necessarily imply equally good performance on another. To avoid laborious evaluation across multiple extrinsic tests a number of intrinsic tasks are used. Ideally they would predict how a model performs in downstream applications. However, it has been shown that intrinsic and extrinsic scores do not always correlate (Tsvetkov et al., 2015; Schnabel et al., 2015).

This study discusses the methodology behind several existing intrinsic evaluations for word embeddings, showing that their chief premise is "interpretability" of a model as a measure of its quality. This approach has methodological issues, and it also ignores the unique feature of word embeddings - their ability to represent fluidity and fuzziness of meaning that is unattainable by traditional linguistic analysis. We argue for a shift from absolute ratings of word embeddings towards more exploratory evaluations that would aim not for generic scores, but for identification of strengths and weaknesses of embeddings, thus providing better predictions about their performance in downstream tasks.

2 Existing Intrinsic Evaluations

2.1 Word Similarity and Relatedness Tests

The term "semantic relatedness" is used to refer to any kind of semantic relation between words. The degree of semantic relatedness reflects the degree to which two words share attributes (Turney et al., 2010, p. 149). Similarity is defined by Turney as co-hyponymy (e.g. *car* and *bicycle*), whereas Hill et al. (2015) define it as "the similarity relation is exemplified by pairs of synonyms; words with identical referents" (e.g. *mug* and *cup*).

The widely used relatedness test sets include WordSim-353 (Finkelstein et al., 2002) and MEN (Bruni et al., 2014)[1]. The former contains 353 word pairs, and the latter - 3,000 word pairs with their relatedness ratings by human annotators. On the other hand, SimLex999 (Hill et al., 2015) specializes on semantic similarity.

The task in cases of both semantic relatedness and semantic similarity is to rate the semantic proximity of two words, usually with the cosine similarity metric. The "best" model is the one

[1]Note that both of these sets also include semantically similar words as a subset of semantic relatedness, e.g. "*cathedral, church*" in MEN and "*football, soccer*" in WordSim.

Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP, pages 36–42,
Berlin, Germany, August 12, 2016.

that comes closest to the ratings of human annotators. Therefore these tests directly assesses interpretability of the model's output - to what extent it mimics human judgments of semantic relations.

The immediate problem with the similarity and relatedness tests is that distributional similarity conflates not only semantic similarity and relatedness, but also morphological relations and simply collocations, and it is not clear whether a model should generally score higher for preferring either of them. Specializing on one of these relations (Kiela et al., 2015) is certainly useful for specific downstream applications, but it would not make a word embedding either generally "good" or universally applicable.

Another concern is, traditionally, the (un)reliability of human linguistic judgements, which are subject to over 50 potential linguistic, psychological, and social confounds (Schutze, 1996). With Amazon Mechanical Turk, typically used to collect ratings, it is impossible to ensure that the participants are native speakers, to get accurate timing, or to control the environment in which they provide responses. Inter-annotator agreement provides an estimate of uniformity of the data, but, if there is a general problem, we would not detect it.

Semantic relatedness is particularly confusing to rate. Consider WordSim scores for hyponymy and hypernymy: "*money, dollar*" (8.42) vs "*tiger, mammal*" (6.85). There is no theoretical ground for rating either semantic relation higher; subjects are likely to rank based on frequency, prototypicality, and speed of association, and not "semantic relatedness" *per se*.

It is also worth mentioning that word embeddings vary in the amount of frequency information that they encode, and frequency can confound estimates of relatedness (Schnabel et al., 2015; Wilson and Schakel, 2015). Thus, depending on the embedding, results of tests such as WordSim need to be considered in the context of the corpus.

2.2 Comparative Intrinsic Evaluation

The comparative intrinsic evaluation for word embeddings was introduced by Schnabel et al. (2015). Several models are trained on the same corpus, and polled for the nearest neighbors of words from a test set. For each word, human raters choose the most "similar" answer, and the model that gets the most votes is deemed the best.

The advantage of this method is the possibility to compare first, second, etc. nearest neighbors in different models. However, it inherits the problem with human interpretation of distributional similarity, which we discussed above. Consider the examples[2] in table 1:

	Target word	GloVe	SVD
1	phone	telephone	mobile
2	coffee	tea	drinks
3	grammar	vocabulary	grammatical
4	cohesiveness	cohesion	inclusiveness

Table 1: Examples of nearest neighbors in GloVe and SVD

Subjects asked to choose the most "similar" word would presumably prefer synonyms (word 1 in table 1), if any were present (thus the "best" model would be the one favoring similarity over relatedness). They would easily exclude the clearly unrelated words (word 4 for SVD model). But they would provide less reliable feedback on "related" options, where the choice would be between different semantic relations (words 2,3). Many answers would be subjective, if not random, and likely to reflect frequency, speed of association, and possibly the order of presentation of words - rather than purely semantic factors that we are trying to evaluate.

2.3 "Coherence" of Semantic Space

Schnabel et al. (2015) also suggested that a "good" word embedding should have coherent neighborhoods for each word vector. The test they proposed consists in choosing two nearest neighbors of a test word, and adding a random word. A human rater should be able to identify the "intruder". For example, in our GloVe the nearest neighbors of *true* are *indeed* and *fact*; they are more semantically related to each other than to a random word *taxi*.

This test still relies on human interpretation, but it is more likely to produce reliable results than the methods discussed above. However, to apply it on

[2]Unless specified otherwise, the examples cited in this study are derived from 2 word embeddings: GloVe (Pennington et al., 2014) and SVD, trained at 300 dimensions, window size 10. GloVe parameters: 100 iterations, x_{max}= 100, a = 3/4. The SVD (Singular Vector Decomposition) model was built with Pointwise Mutual Information (PMI), $a = 1$, using the co-occurrence extraction kernel by Drozd et al. (2015). The 5B web-corpus combines Wikipedia (1.8B tokens), Araneum Anglicum Maius (1.2B) (Benko, 2014) and ukWaC (2B) (Baroni et al., 2009).

a large scale we would need to establish the size of neighborhoods that should be coherent. This number differs between words (see examples in table 2), and a "better" model should recognize that *beautiful* has more "good" neighbors than *knob*. But it is hard to tell the exact number a priori, and independently of a particular corpus.

beautiful	write	knob
lovely, 0.81	writing, 0.75	knobs, 0.60
gorgeous, 0.77	read, 0.72	gearshift, 0.48
wonderful, 0.64	written, 0.65	toggle, 0.41
magnificent, 0.63	want, 0.64	dials, 0.40
elegant, 0.61	wish, 0.62	dashboard, 0.38

Table 2: Nearest neighbors of *beautiful*, *write* and *knob* in GloVe

2.4 Alignment with Linguistic Features

Tsvetkov et al. (2015) proposed the QVec system that evaluates how well the dimensions of a word embedding can be aligned with dimensions of "linguistic" vectors (constructed from a semantically annotated corpus)[3]. This system does not involve human raters, but it still focuses on the "interpretability", as any linguistic categories are themselves a product of a certain theoretical interpretation of the language system.

The core assumption of QVec is that dimensions of word embeddings correspond to linguistic features (in this case, 41 supersenses of WordNet (Miller and Fellbaum, 1998) such as *food* or *motion*). Each linguistic feature can be mapped onto several dimensions of the word embedding, but each dimension of the word embedding can be mapped onto at most one linguistic feature. This assumption could be challenged: it is not clear why one dimension could not encode several linguistic features, or even that a certain cluster or pattern of dimensions could not correspond to one or several linguistic features.

Crucially, the authors report that the correlation of QVec with performance on different tasks varies with vector dimensionality (0.32 for 50 dimensions, 0.78 for 300 and 0.60 for 1000 on the sentiment analysis task). Such variation could be explained by the intuition that in smaller word embeddings dimensions have to be multi-functional, and in larger embeddings more complex patterns of correspondence could be expected to occur. And increasingly complex patterns are likely to make decreasing amount of sense to humans.

[3]See also (Faruqui et al., 2015) for decomposition of word embeddings into sparse vectors to increase their correspondence to linguistic features. Such vectors are reported to be more "interpretable" to human annotators in the word intrusion task.

3 General Methodological Concerns

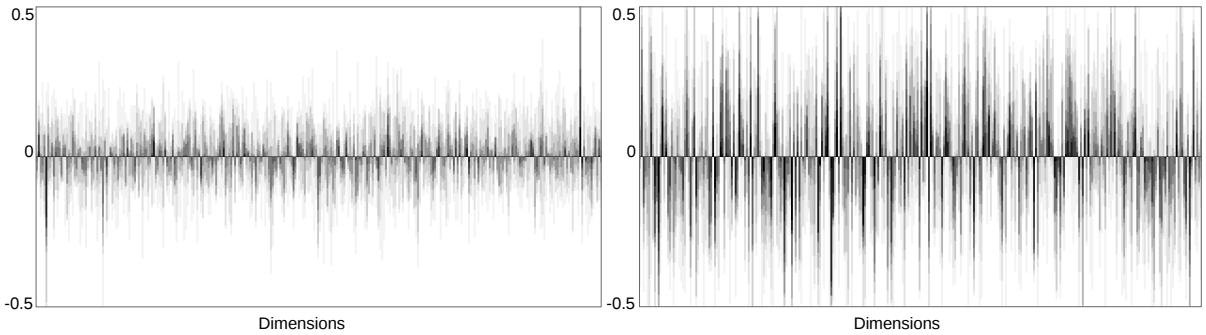
3.1 Do Dimensions Have to Be Interpretable?

Although both people and word embeddings acquire the meanings of words from context, there are many important differences between human semantic knowledge and what can be expected from word embeddings. The latter depend on corpora that are static, noisy, and small. Co-occurrence frequencies do not mirror the frequencies of events that give rise to natural language semantics (e.g. "*dog bites man*" is less likely to be mentioned than "*man bites dog*") (Erk, 2016).

Thus even the most perfect word embedding is unlikely to have exactly the same "concepts" as us, or that their structure would mirror the categories of some linguistic theory. QVec proves that to some extent the dimensions of the vector space are indeed interpretable, but the point we would like to make is this: by focusing on the structures that we expect the word embeddings to have, we might be missing the structures that they actually have.

Figure 1 compares the overlap of dimensions for 10 random words and 10 co-hyponyms in 300-dimensional GloVe vectors (darker dimensions indicate overlap between more words in the sample). It is clear that there are hundreds of features relevant for felines. We could hypothesize about them ("animal"? "nounhood"? "catness"?), but clearly this embedding has more "feline" features thanwhat we could find in dictionaries or elicit from human subjects. Some of such features might not even be in our conceptual inventory. Perhaps there is a dimension or a group of dimensions created by the co-occurrences with words like *jump*, *stretch*, *hunt*, and *purr* - some "feline behavior" category that we would not find in any linguistic resource.

Distributional models are gradient by nature. This makes them less interpretable, but also more similar to connectionist cognitive models (Lenci, 2008). We do not know to what extent word embeddings are cognitively plausible, but they do offer a new way to represent meaning that goes beyond symbolic approaches. We would be missing the point if we were only seeking features that we



10 random words: *emergency, bluff, buffet, horn, human, like, american, pretend, tongue, green*

10 felines: *cat, lion, tiger, leopard, cougar, cheetah, lynx, bobcat, panther, puma*

Figure 1: Heatmap histogram of 10 random words and 10 co-hyponyms in GloVe

know from traditional linguistics.

3.2 Polysemy: the Elephant in the Room

Another general problem with all evaluations discussed above stems from the (lack of) treatment of polysemy in word-level word embeddings. Do we expect the vector for *apple* to be closer to *computer* or to *pear*? The cosine similarity-based tests choose only the "strongest" sense of a word in a given corpus. Therefore the accuracy of the current intrinsic evaluation methods also depends on whether the relations in the test word pairs match the distribution of senses of these words in a particular corpus. "*Apple, pear*" could be rated low, and "*apple, computer*" - high, but preference for either pair would say nothing about quality of the word embedding itself.

One way to deal with this problem is to exclude ambiguous words from tests, as it is done in BLESS; but this would be hard to guarantee for all corpora, it would significantly limit the tests (as more frequent words tend to be more polysemous), and it would avoid the issue rather than deal with it. Alternatively, we could attempt word sense disambiguation (Neelakantan et al., 2014; Bartunov et al., 2015); but the accuracy would be hard to guarantee, and we would need to provide the mapping from the word senses in the test to the word senses in the corpus.

The alternative is to embrace ambiguity as an intrinsic characteristic of word embeddings. We are looking for interpretable dimensions because we are used to discrete linguistic features, and similarly we are trying to bring meaning representations in word embeddings down to neat lists of word senses in dictionaries that we are used to. But anyone who has done lexicographic work knows that dictionaries are only an abstraction, never complete or free of inconsistencies and subjectivity. The distributional approach offers us a novel way to capture the full continuum of meaning (Erk, 2009). From this perspective, the problem with polysemy in tests for word embeddings is not the polysemy itself, but the fact that we are ignoring it with out-of-context test words and cosine similarity.

4 Back to the Drawing Board

4.1 What We Should Start Thinking About

To sum up, all intrinsic evaluations of word embeddings discussed above are based on the idea of interpretability by humans, and suffer from the problem of word ambiguity. We argue that both problems stem from the underlying methodological principle - the attempt to transfer the traditional lexicographic model of discrete word senses and linguistic features onto the continuous semantic space.

The reason that this methodology is so widespread is that linguistics does not yet offer an alternative, and finding one would require a lot of (collaborative) work by both theoretical and computational linguists. We will need to think of answers to some very basic questions. For example, how granular do we want our semantics to be? (individual word senses? lexical groups?) Should embeddings aim at separating word groups as neatly as possible, or rather at blending them by giving more weight to cases that would puzzle human annotators? The former would be easier to work with from the point of view of downstream applications; the latter would arguably provide a truer model of language for the linguists.

With respect to "interpretability" of word em-

beddings, the biggest question is the nature of those potentially non-interpretable dimensions. We can continue ignoring them and work only with the ones we can understand (which could prove to be enough for certain applications). The alternative is to accept that from now on we will not really understand our semantics, and delegate the interpretation to machine learning algorithms.

4.2 What Can We Do Right Now?

The above discussion does not yet offer any alternatives to current evaluations of word embeddings, but it does offer some insights about their interpretation. Things that we can learn from existing tests include:

- the degree to which a word embedding encodes frequency information, and is likely to be biased by it (Schnabel et al., 2015; Wilson and Schakel, 2015);
- the richness of representations for rare words (Wartena, 2014);
- performance on different size of corpora (while more data is mostly better, we also need "good" word embeddings for low-resource languages);
- specialization for a particular type of relation in distributional similarity, if any.

The last option is explored in such test sets as BLESS (Baroni and Lenci, 2011) and EVALution (Santus et al., 2015). They include pairs of words with different kinds of relations, such as synonymy and meronymy, but no annotator ratings. The word embeddings are queried on similarity between these pairs of words. The distribution of similarity ratings across different relations shows what linguistic relations are "favored" by the given embedding. This approach can be fruitfully extended to other types of linguistic relations, such as derivational morphology and frame relations.

Ideally, evaluations of a new model would also include publishing results of systematic tests for different parameters (Levy et al., 2015; Lai et al., 2015) and types of context (Melamud et al., 2016), as well as different types of linguistic relations (Gladkova et al., 2016). This kind of data is often viewed as something simply to be used for choosing a model for a particular task - but it does also offer insights into its nature, and could help us understand the deeper properties of word embeddings, which could eventually lead to new types of tests.

None of these above-mentioned characteristics of word embeddings provides a one-number answer about how "good" a model is. But we can take a more exploratory approach, identifying the properties of a model rather than aiming to establish its superiority to others.

Lastly, when evaluating word embeddings we should not forget that the result of any evaluation is down to not only the embedding itself, but also the test, the corpus, and the method of identifying particular relations. Thus we cannot interpret, e.g., a low score on analogy test as evidence that a given model does not *contain* some linguistic feature: all it means is that we could not *detect* it with a given method, and perhaps a different method would work better (Drozd and Matsuoka, 2016).

5 Conclusion

This paper discusses the current methods of intrinsic evaluation of word embeddings. We show that they rely on "interpretability" of the model's output or structure, and we argue that this might not be the best approach, as it ignores the key features of distributional semantics, and does not always yield good predictions for how a word embedding would perform on a downstream application. We suggest focusing not on absolute ratings of abstract "quality" of embeddings, but on exploration of their characteristics.

We hope to draw attention of both computational and theoretical linguists to the need of working together on new models of language that would help us make better sense, and better use, of word embeddings.

References

Marco Baroni and Alessandro Lenci. 2011. How we BLESSed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, GEMS '11, pages 1–10. Association for Computational Linguistics.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. 2015. Breaking sticks and ambiguities with adaptive skip-gram. *arXiv:1502.07257*.

Vladimír Benko. 2014. Aranea: Yet another family of (comparable) web corpora. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, speech, and dialogue: 17th international conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings*, LNCS 8655, pages 257–264. Springer.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal Distributional Semantics. *JAIR*, 49(1-47).

Danqi Chen and Christopher D. Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP) 2014*, pages 740–750.

Yun-Nung Chen, William Yang Wang, and Alexander I. Rudnicky. 2014. Leveraging frame semantics and distributional semantics for unsupervised semantic slot induction in spoken dialogue systems. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pages 584–589. IEEE.

Aleksandr Drozd and Satoshi Matsuoka. 2016. Linguistic regularities from multiple samples. Technical Report C-283, Department of Mathematical and Computing Sciences, Tokyo Institute of Technology.

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. 2015. Python, performance, and natural language processing. In *Proceedings of the 5th Workshop on Python for High-Performance and Scientific Computing*, PyHPC '15, pages 1:1–1:10, New York, NY, USA. ACM.

Katrin Erk. 2009. Supporting inferences in semantic space: representing words as regions. In *Proceedings of the Eighth International Conference on Computational Semantics*, IWCS-8 '09, pages 104–115. Association for Computational Linguistics.

Katrin Erk. 2016. What do you know about an alligator when you know the company it keeps? *Semantics and Pragmatics*, 9(17):1–63, April.

Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah Smith. 2015. Sparse overcomplete word vector representations. *arXiv:1506.02004*.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. In *ACM Transactions on Information Systems,*, volume 20(1), pages 116–131. ACM.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of NAACL-HLT 2016*, pages 47–54. Association for Computational Linguistics.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Revisiting embedding features for simple semi-supervised learning. In *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP) 2014*, pages 110–120.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Douwe Kiela, Felix Hill, and Stephen Clark. 2015. Specializing word embeddings for similarity or relatedness. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP))*, pages 2044–2048. Association for Computational Linguistics.

Siwei Lai, Kang Liu, Liheng Xu, and Jun Zhao. 2015. How to generate a good word embedding? *arXiv:1507.05523*.

Alessandro Lenci. 2008. Distributional semantics in linguistic and cognitive research. *From context to meaning: Distributional models of the lexicon in linguistics and cognitive science, special issue of the Italian Journal of Linguistics*, 20(1):1–31.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. In *Transactions of the Association for Computational Linguistics*, volume 3, pages 211–225.

Oren Melamud, David McClosky, Siddharth Patwardhan, and Mohit Bansal. 2016. The role of context types and dimensionality in learning word embeddings. *arXiv:1601.00893*.

George Miller and Christiane Fellbaum. 1998. *Wordnet: An electronic lexical database*. MIT Press: Cambridge.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, pages 1059–1069. Association for Computational Linguistics.

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

Enrico Santus, Frances Yung, Alessandro Lenci, and Chu-Ren Huang. 2015. EVALution 1.0: an evolving semantic dataset for training and evaluation of distributional semantic models. In *Proceedings of the 4th Workshop on Linked Data in Linguistics (LDL-2015)*, pages 64–69. Association for Computational Linguistics.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015), Lisbon, Portugal*, pages 298–307. Association for Computational Linguistics.

Carson T. Schutze. 1996. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. University of Chicago Press.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. 2015. Evaluation of word vector representations by subspace alignment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2049–2054. Association for Computational Linguistics.

Peter D. Turney, Patrick Pantel, and others. 2010. From frequency to meaning: Vector space models of semantics. *JAIR*, 37(1):141–188.

Christian Wartena. 2014. On the effect of word frequency on distributional similarity. In *Proceedings of the 12th edition of the KONVENS conference - Hildesheim*, volume 1, pages 1–10.

Benjamin J. Wilson and Adriaan MJ Schakel. 2015. Controlled experiments for word embeddings. *arXiv:1510.02675*.