An Improved Crowdsourcing Based Evaluation Technique for Word Embedding Methods

Farhana Ferdousi Liza
School of Computing
University of Kent
Canterbury, CT2 7NY, UK
fl207@kent.ac.uk

Marek Grześ
School of Computing
University of Kent
Canterbury, CT2 7NY, UK
m.grzes@kent.ac.uk

Abstract

In this proposal track paper, we have presented a crowdsourcing-based word embedding evaluation technique that will be more reliable and linguistically justified. The method is designed for intrinsic evaluation and extends the approach proposed in (Schnabel et al., 2015). Our improved evaluation technique captures word relatedness based on the word context.

1 Introduction

The semantic relatedness between words can be ambiguous if the context of the word is not known (Patwardhan et al., 2003), and word sense disambiguation is the process of assigning a meaning to a polysemous word based on its context. The context defines linguistic and corresponding factual real world knowledge which provides a difference between word's sense and its reference. The sense of a word concerns one of the meanings of a word in a particular language. Reference is used to deal with the relationship between a language and the real world knowledge about an object or entity. The context of a word can be understood through a sentence, and thus understanding a word in a sentential context works as ambiguity resolution (Faust and Chiarello, 1998).

The vector space representation of words (embeddings) keeps related words nearby in the vector space. The word relatedness is usually measured through synonyms, but synonyms can differ in at least one semantic feature. The feature can be 'denotative', referring to some actual, real world difference in the object the language is dealing with, such as, walk, lumber, stroll, meander, lurch, stagger. The feature can be 'connotative', referring to how the user feels about the object rather than any real difference in the object itself, such as, die, pass away, give up the ghost, kick the bucket, croak. Absolute synonyms are usually rare in a language. For example: sofa and couch are nearly absolute synonyms, however based on the context, they have different meaning in at least one way, such as, couch potato, because there is no word sense available for sofa potato (Vajda, 2001).

Crowdsourcing (Ambati et al., 2010; Callison-Burch, 2009), which allows employing people worldwide to perform short tasks via online platforms, can be an effective tool for performing evaluation in a time and cost-effective way (Ambati, 2012). In (Schnabel et al., 2015), crowdsourcing-based evaluation was proposed for synonyms or a word relatedness task where six word embedding techniques were evaluated. The crowdsourcing-based intrinsic evaluation which tests embeddings for semantic relationship between words focuses on a direct comparison of word embeddings with respect to individual queries. Although the method is promising for evaluating different word embeddings, it has some shortcomings. Specifically, it does not explicitly consider word context. As the approach relies on human interpretation of words, it is important to take into account how humans interpret or understand the meaning of a word. Humans usually understand semantic relatedness between words based on the context. Thus, if the approach is based only on the word without its context, it will be difficult for humans to understand the meaning of a particular word, and it could result in word sense ambiguity (WSA).

In this paper, we show what are the consequences of the lack of the word context in (Schnabel et al., 2015), and we discuss how to address the resulting challenge. Specifically, we add a sentential context to mitigate word sense ambiguity, and this extension leads to an improved evaluation technique that explicitly accounts for multiple senses of a word.

2 Crowdsourcing Evaluation

2.1 Details of the Method

The method in (Schnabel et al., 2015) started by creating a *query inventory* which is a pre-selected set of query terms and semantically related target words. The query inventory consists of 100 query terms that balance frequency, part of speech (POS), and concreteness. The query terms were selected from 10 out of 45 broad categories from WordNet (Miller, 1995). Then, 10 random words with one adjective, one adverb, four nouns, and four verbs were selected based on concrete concepts from each category. Among the 10 words, 3 words were rare with the property that the number of their occurrences in the training corpus—Wikipedia dump (2008-03-01)—is smaller than 2500.

For each of those 100 query terms in the inventory, the nearest neighbours at ranks $k \in \{1, 5, 50\}$ for the six embeddings from CBOW (Mikolov et al., 2013), Glove (Pennington et al., 2014), TSCCA (Dhillon et al., 2012), C&W (Collobert et al., 2011), H-PCA (Lebret and Lebret, 2013), and Random Projection (Li et al., 2006) were retrieved. Then, for each k, the query word along with the six words corresponding to the embeddings described above were presented to human testers (Turkers) from Amazon Mechanical Turk (MTurk) for evaluation. Each Turker was requested to evaluate between 20 and 50 items per task, where an item corresponds to the query term and a set of 6 retrieved nearest neighbour words from each of the six embeddings. The Turkers' were then asked to select one of the six words that is the closest synonym to the query word according to their perception. For the selected 100 query words and 3 ranks (k), there were a total of 300 terms on which Turkers' perception-based choices were used for evaluating the embedding techniques. The comparison of embeddings was done by averaging the win ratio, where the win ratio was how many times the Turker chose a particular embedding divided by the number of total ratings for the corresponding query word.

2.2 Shortcomings of the Method

A word relatedness evaluation task for word embeddings is challenging due to ambiguity inherent in word sense and corresponding reference. Although the experiments in (Schnabel et al., 2015) incorporated participants with adequate knowledge of English, the ambiguity is inherent in the language. This means that evaluations that ignore the context may have impact on the evaluation result. Also, the evaluated word embedding techniques in (Schnabel et al., 2015)—except TSCCA (Dhillon et al., 2015)—generate one vector for each word, and that makes comparisons between two related words from two embedding techniques difficult. For example, the word 'bank' may be embedded by CBOW as a noun in the context of 'he cashed a cheque at the bank' where the related word according to nearest neighbours would be 'financial' or 'finance' whereas the TSCCA might embed the same 'bank' as a noun but in the context of 'they pulled the canoe up on the bank' where related word according to nearest neighbours would be 'slope' or 'incline'. Although all the embedding techniques have been trained with the same corpus, different techniques may encode different explanatory factors of variation present in the data (Gao et al., 2014), and using one embedding vector per word cannot capture the different meanings (Huang et al., 2012), and as a result, not all senses will be conflated into one representation.

If the query word 'bank' is presented to a user with 'financial' and 'incline' as related words, and a user is asked which one is more likely to be a related word, then the user has to choose one word, but she does not know the context. Therefore, if 100 people were asked to evaluate the query word, and 50 persons voted for 'financial' and 50 persons voted for 'incline' to be a related word, then both CBOW and TSCCA have the same score. However, this judgement would be inaccurate as CBOW can embed one vector per word whereas TSCCA can embed multiple vectors for each word. Thus user's choice of a related word does not have sufficient impact on the quality evaluation of the embedding techniques. Note that the word 'bank', as a noun, has 10 senses in WordNet.

Before we introduce our extensions in the next section, we investigate how (Schnabel et al., 2015) accommodates word sense ambiguity. The Turker is presented with a query word and several related words to choose from. If the options presented to the Turker are from different contexts, the Turker has to choose from several correct senses. The Turker could be instructed that multiple senses can be encountered during the experiment, and one of

the two alternative solutions could be considered:

1. **Biased** Select the sense that is most likely according to your knowledge of the language
2. **Uniform sampling** Select one sense randomly giving the same preference to all options

The first approach would be more appropriate because senses that are more common would be given higher priority. The second option would be hard to implement in practice because it is not clear if random sampling could be achieved, but this option will be useful to show connections with our method. Certainly, even if the Turker can sample according to a uniform probability, the real samples would depend on which senses contained in the corpus were captured by various word embedding techniques. Overall, using the above options, one could argue that the method accommodates different senses because the evaluation measures how well the word embedding methods recover the sense selection strategy of the user. The biased method would be desirable because it would focus on the most frequent senses, but one should note that this would depend on the subjective judgement of the user and her knowledge.

3 Proposed Extensions

Recent efforts on multiple embeddings for words (Neelakantan et al., 2015; Reisinger and Mooney, 2010) require a more sophisticated evaluation and further motivate our ideas. There are existing works, such as (Song, 2016; Iacobacci et al., 2015), where the sense embedding was proposed as a remedy for the current word embedding limitation on ubiquitous polysemous words, and the method learns a vector for each sense of a word. For words with multiple meanings, it is important to see how many senses a word embedding technique can represent through multiple vectors. To achieve such an evaluation, we have first extended the work of (Schnabel et al., 2015) to include sentential context to avoid word sense ambiguity faced by a human tester. In our method, every query word is accompanied by a context sentence. We then extended the method further so that it is more suitable to evaluate embedding techniques designed for polysemous words with regard to their ability to embed diverse senses.

3.1 First Extension

Our chief idea is to extend the work of (Schnabel et al., 2015) by adding a context sentence for each query term. Using a context sentence for resolving word sense ambiguity is not a new concept, and it has been used by numerous researchers, such as (Melamud et al., 2015; Huang et al., 2012; Stetina et al., 1998; Biemann, 2013). In particular, human judgement based approaches, such as (Huang et al., 2012), have used the sentential context to determine the similarity between two words, and (Biemann, 2013) used sentential context for lexical substitution realising the importance of the word interpretation in the context for crowdsourcing-based evaluations.

Due to limited and potentially insufficient embedded vocabulary used to identify a related sense of the query term, we are also proposing to provide another option of 'None of the above' along with the six words. In fact, (Schnabel et al., 2015) have already considered 'I don't know the meaning of one (or several) of the words'; however, when the context is in place, there may be a situation when none of the embeddings make a good match for the query term, and in that case 'None of the above' is more appropriate. In this way, the user's response will be more justified, and a more reliable evaluation score will be retrieved. Our proposal is based on an observation that human reasoning about a word is based on the context, and in crowdsourcing evaluations, we use a human to interpret the meaning; and based on their judgement, we evaluate embedding techniques. So the human should be presented with the examples in the manner that is consistent with what humans see in real-life.

3.2 Second Extension

In our first extension above, every query word is presented in a context. In order to implement a multi-sense evaluation, every query word is presented in several contexts where contexts represent different senses. The number (p) of the contexts presented, where $p \geq 1$, will depend on the number and frequency of available senses for a particular query word. Note that p contexts for the query word are presented in every round, and the Turker has more senses to choose from when word embeddings encode multiple senses per word.

3.3 Example

The true, related words are those that are retrieved from the embedding techniques using the nearest neighbour algorithm, for example. Below, we show an example word 'bar' together with its context; the context is extracted from WordNet.

Query Word: **Bar**, [Context Sentence: He drowned his sorrows in whiskey at the bar.], {True Related Words: barroom, bar, saloon, ginmill, taproom}

To extend the evaluation for multi-sense embedding capabilities of the embedding techniques, we will extend the example setting above by adding multiple test cases for each query word representing different senses. Note that this is not needed in (Schnabel et al., 2015) where query words are not annotated. In the above example, only one test case per query word was presented. However, for the query word 'Bar' as a noun, there are 15 senses available in WordNet 3.0, and 23 senses available in 2012 version of Wikipedia (Dandala et al., 2013a). For the second extension, the human evaluator will be presented with p context sentences representing p different senses. The criteria for selecting senses, and the corresponding context sentences will be discussed in the next section.

3.4 Context Generation

In every iteration, every word embedding method will return its best match for the query term. Our method will need to determine a context (i.e. an appropriate sentence for the given word). We call this process context generation, and this section introduces two approaches that can be used to implement it.

3.4.1 Informed Matching

In this informed approach, our assumption is that the senses selected for the query word should exist in the training corpus. Below we explain how to implement this feature.

Matching Frequent Senses In this approach, the goal is to use the most frequent senses from WordNet. In this way, we can take into account the frequency of senses embedded in WordNet. For every query word, the most frequent n, where $n \geq 1$, word senses will be selected from WordNet. Note that we have to select only those senses that exist in our training corpus which is Wikipedia in this case. The mapping of the senses between Wikipedia and WordNet will be implemented using a method similar to (Mihalcea, 2007, Section 3.1). In the final step of their method, the labels (Wikipedia senses) are manually (i.e. they are performed by a human) mapped to WordNet senses. An alternative approach would be automated mapping introduced in (Fernando and Stevenson, 2012), which does not require human intervention. One could argue that the manual mapping would be more accurate because of the incorporation of the human judgement, however, this is expensive and time consuming. As the overlapping, most frequent senses from the Wikipedia and WordNet will be chosen, the correct senses corresponding to the embedded word can be selected by Turkers as long as the word embedding methods are accurate. Since our method presents n senses per run, it is more likely that one or more of the chosen senses were embedded by the embedding techniques. Note that senses in WordNet are generally ordered from the most frequent to the least frequent. WordNet sense frequencies come from the SemCor (Miller et al., 1993) sense-tagged corpus which means that WordNet frequencies are well justified, and they are based on data. The example sentence corresponding to the chosen sense will be taken as a context sentence. As WordNet was annotated by humans, we assume that the context sentences are correct for a particular sense.

Matching Rare Senses In (Vossen et al., 2013), the authors argue that current *sense-tagged* corpora have insufficient support for rare senses and contexts and, as a result, they may not be sufficient for word-sense-disambiguation. For example, WordNet 3.0 has 15 senses for the word 'bar' as a noun, whereas 2012 version of Wikipedia has 23 senses (Dandala et al., 2013a) for this word. As a remedy for this issue, we propose another way to generate contexts where we utilise m, where $m \geq 1$, randomly selected senses from the training corpus (Wikipedia in our case). Note that this section applies to the situation where none of the rare senses exist in WordNet. Since Wikipedia does not contain frequencies for senses, sampling has to be according to a uniform distribution. Overall, Wikipedia can be used as a training corpus for the embedding methods and also for sense annotation.

In (Mihalcea, 2007), the authors showed that links in Wikipedia articles are appropriate for representing a sense. When Wikipedia will be used for selecting rare senses, the context sentence will be retrieved using a similar method to (Mihalcea, 2007, Section 3.1). Specifically, in the final step of the mapping method of (Mihalcea, 2007, Section 3.1), the labels (Wikipedia senses) were mapped to WordNet senses. However, this time we are interested in the word senses that are not available in WordNet; as a result, we will map the selected senses from Wikipedia to the appropri-

ate subsenses in the Oxford Dictionary of English (ODE) (Soanes and Stevenson, 2003). Note that ODE provides a hierarchical structure of senses, and each polysemous sense is divided into a core sense and a set of subsenses (Navigli, 2006). We will follow an approach similar to (Navigli, 2006) where WordNet sense was semantically mapped to the ODE core senses. They mapped to the core senses because they were interested in the coarse-grained sense mapping to resolve granularity inherent in WordNet. In our case, we will do semantic mapping between Wikipedia senses (piped link or simple link) and ODE subsenses, instead of mapping the WordNet sense to the ODE core senses. Then, corresponding context sentences will be selected from Wikipedia or ODE.

Overall, when the corresponding context sentence for a query term is not available in WordNet, the context sentence can be retrieved from Wikipedia (Mihalcea, 2007; Dandala et al., 2013b) or ODE using the method described above.

3.4.2 Random Matching

The informed method described above requires either manual matching by humans (which are time consuming and expensive) or an automated matching which may be inaccurate. An alternative approach is to sample senses randomly from WordNet ignoring senses contained in the training corpus. The sampling distribution should be based on frequencies of senses. In this case, 'None of the above' option will be used whenever none of the embedded words are related to the query word according to the presented context. If we consider a large number of Turkers' evaluations, the evaluation will still give the performance score reflecting the true performance score of the embedding technique. However, this will be more costly because more Turkers will be required.

3.5 Merit of our Extensions

At the end of Sec. 2.2, we explained how word sense ambiguity is accommodated in (Schnabel et al., 2015). We argued that their evaluation was in expectation with respect to subjective preferences of the Turkers. Additionally, when the context is not provided, the Turkers may even forget about common senses of the query word. In our proposal, we argue that query words should be presented in an appropriate context. Similar to Sec. 2.2, we can distinguish two ways in which we can apply our method:

1. **Informed sampling** Sample senses according to their frequency in WordNet
2. **Uniform sampling** Sample senses according to a uniform probability distribution if no frequency data is available (e.g. Wikipedia)

We can now draw a parallel with alternative ways that Turkers may apply to solve the word sense ambiguity problem. In particular, under certain conditions (i.e. when word embeddings don't use sense frequency information), the uniform sampling option in our method would be equivalent with the uniform sampling method in Sec. 2.2. This means that asking the Turkers to select senses randomly according to a uniform probability distribution is the same as sampling contexts according to a uniform distribution. The two approaches differ, however, when non-uniform, informed probability distributions are used. Informed sampling in our approach is based on WordNet whose sense frequencies are based on data-driven research. This means that the overall evaluation would be based on real frequencies coming from the data instead of subjective and idiosyncratic judgements by the Turkers. This probabilistic argument provides another justification for our approach.

4 Conclusion

In this paper, a crowdsourcing-based word embedding evaluation technique of (Schnabel et al., 2015) was extended to provide data-driven treatment of word sense ambiguity. The method of (Schnabel et al., 2015) relies on user's subjective and knowledge dependent ability to select 'preferred' meanings whereas our method would deal with this problem selecting explicit contexts for words. The selection is according to the real frequencies of meanings computed from data. With this data-driven feature, our method could be more appropriate to evaluate both methods that produce one embedding per *word* as well as methods that produce one embedding per *word sense*. Our method would provide scores that accommodate word sense frequencies in the real use of the language. Here, we assume that word embeddings should recover the most frequent senses with higher priority.

Acknowledgement We thank the anonymous reviewers for stimulating criticism and for pointing out important references. We also thank Omer Levy for being a supportive workshop organiser.

References

Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. 2010. Active learning and crowd-sourcing for machine translation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC-2010)*, pages 2169–2174, Valletta, Malta, May. European Languages Resources Association (ELRA). ACL Anthology Identifier: L10-1165.

Vamshi Ambati. 2012. *Active Learning and Crowdsourcing for Machine Translation in Low Resource Scenarios*. Ph.D. thesis, Pittsburgh, PA, USA. AAI3528171.

Chris Biemann. 2013. Creating a system for lexical substitutions from scratch using crowdsourcing. *Language Resources and Evaluation*, 47(1):97–122.

Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 286–295, Singapore, August. Association for Computational Linguistics.

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.

Bharath Dandala, Chris Hokamp, Rada Mihalcea, and Razvan C. Bunescu. 2013a. Sense clustering using wikipedia. In Galia Angelova, Kalina Bontcheva, and Ruslan Mitkov, editors, *Recent Advances in Natural Language Processing, RANLP 2013, 9-11 September, 2013, Hissar, Bulgaria*, pages 164–171. RANLP 2013 Organising Committee / ACL.

Bharath Dandala, Rada Mihalcea, and Razvan Bunescu. 2013b. Word sense disambiguation using wikipedia. In *The People's Web Meets NLP*, pages 241–262. Springer.

Paramveer S. Dhillon, Jordan Rodu, Dean P. Foster, and Lyle H. Ungar. 2012. Two step cca: A new spectral method for estimating vector models of words. In *Proceedings of the 29th International Conference on Machine learning*, ICML'12.

Paramveer S. Dhillon, Dean P. Foster, and Lyle H. Ungar. 2015. Eigenwords: Spectral word embeddings. *Journal of Machine Learning Research*, 16:3035–3078.

Miriam Faust and Christine Chiarello. 1998. Sentence context and lexical ambiguity resolution by the two hemispheres. *Neuropsychologia*, 36(9):827–835.

Samuel Fernando and Mark Stevenson. 2012. Mapping wordnet synsets to wikipedia articles. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Bin Gao, Jiang Bian, and Tie-Yan Liu. 2014. Wordrep: A benchmark for research on learning word representations. *CoRR*, abs/1407.1640.

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 Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 873–882, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. Sensembed: learning sense embeddings for word and relational similarity. In *Proceedings of ACL*, pages 95–105.

Rémi Lebret and Ronan Lebret. 2013. Word emdeddings through hellinger PCA. *CoRR*, abs/1312.5542.

Ping Li, Trevor J. Hastie, and Kenneth W. Church. 2006. Very sparse random projections. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 287–296, New York, NY, USA. ACM.

Oren Melamud, Omer Levy, and Ido Dagan. 2015. A simple word embedding model for lexical substitution. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*.

Rada Mihalcea. 2007. Using wikipedia for automatic word sense disambiguation. In *HLT-NAACL*, pages 196–203.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, HLT '93, pages 303–308, Stroudsburg, PA, USA. Association for Computational Linguistics.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November.

Roberto Navigli. 2006. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 105–112. Association for Computational Linguistics.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2015. Efficient non-parametric estimation of multiple embeddings per word in vector space. *arXiv preprint arXiv:1504.06654*.

Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. 2003. Using measures of semantic relatedness for word sense disambiguation. In *Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing'03, pages 241–257, Berlin, Heidelberg. Springer-Verlag.

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)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Joseph Reisinger and Raymond J Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117. 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*, pages 298–307, Lisbon, Portugal, September. Association for Computational Linguistics.

Catherine Soanes and Angus Stevenson, editors. 2003. *Oxford Dictionary of English.* Cambridge University Press.

Linfeng Song. 2016. Word embeddings, sense embeddings and their application to word sense induction. The University of Rochester, April.

Jiri Stetina, Sadao Kurohashi, and Makoto Nagao. 1998. General word sense disambiguation method based on a full sentential context. In *In Usage of WordNet in Natural Language Processing, Proceedings of COLING-ACL Workshop*.

Edward Vajda. 2001. Semantics. Webpage for course material of Linguistics 201:INTRODUCTION TO LINGUISTICS.

Piek Vossen, Rubn Izquierdo, and Attila Grg. 2013. Dutchsemcor: in quest of the ideal sense-tagged corpus. In Galia Angelova, Kalina Bontcheva, and Ruslan Mitkov, editors, *RANLP*, pages 710–718. RANLP 2013 Organising Committee / ACL.