Supervised and unsupervised approaches to measuring usage similarity

Milton King and **Paul Cook**
Faculty of Computer Science, University of New Brunswick
Fredericton, NB E3B 5A3, Canada
milton.king@unb.ca, paul.cook@unb.ca

Abstract

Usage similarity (USim) is an approach to determining word meaning in context that does not rely on a sense inventory. Instead, pairs of usages of a target lemma are rated on a scale. In this paper we propose unsupervised approaches to USim based on embeddings for words, contexts, and sentences, and achieve state-of-the-art results over two USim datasets. We further consider supervised approaches to USim, and find that although they outperform unsupervised approaches, they are unable to generalize to lemmas that are unseen in the training data.

1 Usage similarity

Word senses are not discrete. In many cases, for a given instance of a word, multiple senses from a sense inventory are applicable, and to varying degrees (Erk et al., 2009). For example, consider the usage of *wait* in the following sentence taken from Jurgens and Klapaftis (2013):

1. *And is now the time to say I can hardly **wait** for your impending new novel about the Alamo?*

Annotators judged the WordNet (Fellbaum, 1998) senses glossed as 'stay in one place and anticipate or expect something' and 'look forward to the probable occurrence of', to have applicability ratings of 4 out of 5, and 2 out of 5, respectively, for this usage of *wait*. Moreover, Erk et al. (2009) also showed that this issue cannot be addressed simply by choosing a coarser-grained sense inventory. That a clear line cannot be drawn between the various senses of a word has been observed as far back as Johnson (1755). Some have gone so far as to doubt the existence of word senses (Kilgarriff, 1997).

Sense inventories also suffer from a lack of coverage. New words regularly come into usage, as do new senses for established words. Furthermore, domain-specific senses are often not included in general-purpose sense inventories. This issue of coverage is particularly relevant for social media text, which contains a higher rate of out-of-vocabulary words than more-conventional text types (Baldwin et al., 2013).

These issues pose problems for natural language processing tasks such as word sense disambiguation and induction, which rely on, and seek to induce, respectively, sense inventories, and have traditionally assumed that each instance of a word can be assigned one sense.[1] In response to this, alternative approaches to word meaning have been proposed that do not rely on sense inventories. Erk et al. (2009) carried out an annotation task on "usage similarity" (USim), in which the similarity of the meanings of two usages of a given word are rated on a five-point scale.

Lui et al. (2012) proposed the first computational approach to USim. They considered approaches based on topic modelling (Blei et al., 2003), under a wide range of parameter settings, and found that a single topic model for all target lemmas (as opposed to one topic model per target lemma) performed best on the dataset of Erk et al. (2009). Gella et al. (2013) considered USim on Twitter text, noting that this model of word meaning seems particularly well-suited to this text type because of the prevalence of out-of-vocabulary words. Gella et al. (2013) also considered topic modelling-based approaches, achieving their best results using one topic model per

[1] Recent word sense induction systems and evaluations have, however, considered graded senses and multi-sense applicability (Jurgens and Klapaftis, 2013).

Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications, pages 47–52,
Valencia, Spain, April 4 2017.

target word, and a document expansion strategy based on medium frequency hashtags to combat the data sparsity of tweets due to their relatively short length. The methods of Lui et al. (2012) and Gella et al. (2013) are unsupervised; they do not rely on any gold standard USim annotations.

In this paper we propose unsupervised approaches to USim based on embeddings for words (Mikolov et al., 2013; Pennington et al., 2014), contexts (Melamud et al., 2016), and sentences (Kiros et al., 2015), and achieve state-of-the-art results over the USim datasets of both Erk et al. (2009) and Gella et al. (2013). We then consider supervised approaches to USim based on these same methods for forming embeddings, which outperform the unsupervised approaches, but perform poorly on lemmas that are unseen in the training data.

2 USim models

In this section we describe how we represent a target word usage in context, and then how we use these representations in unsupervised and supervised approaches to USim.

2.1 Usage representation

We consider four ways of representing an instance of a target word based on embeddings for words, contexts, and sentences. For word embeddings, we consider word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). In each case we represent a token instance of the target word in a sentence as the average of the word embeddings for the other words occurring in the sentence, excluding stopwords.

Context2vec (Melamud et al., 2016) can be viewed as an extension of word2vec's continuous bag-of-words (CBOW) model. In CBOW, the context of a target word token is represented as the average of the embeddings for words within a fixed window. In contrast, context2vec uses a richer representation based on a bidirectional LSTM capturing the full sentential context of a target word token. During training, context2vec embeds the context of word token instances in the same vector space as word types. As this model explicitly embeds word contexts it seems particularly well-suited to USim.

Kiros et al. (2015) proposed skip-thoughts, a sentence encoder that can be viewed as a sentence-level version of word2vec's skipgram model, i.e., during training, the encoding of a sentence is used to predict surrounding sentences. Kiros et al. (2015) showed that skip-thoughts out-performs previous approaches to measuring sentence-level relatedness. Although our goal is to determine the meaning of a word in context, the meaning of a sentence could be a useful proxy for this.[2]

2.2 Unsupervised approach

In the unsupervised setup, we measure the similarity between two usages of a target word as the cosine similarity between their vector representations, obtained by one of the methods described in Section 2.1. This method does not require gold standard training data.

2.3 Supervised approach

We also consider a supervised approach. For a given pair of token instances of a target word, t_1 and t_2, we first form vectors v_1 and v_2 representing each of the two instances of the target, using one of the approaches in Section 2.1. To represent each pair of instances, we follow the approach of Kiros et al. (2015). We compute the componentwise product, and absolute difference, of v_1 and v_2, and concatenate them. This gives a vector of length $2d$ — where d is the dimensionality of the embeddings used — representing each pair of instances. We then train ridge regression to learn a model to predict the similarity of unseen usage pairs.

3 Materials and methods

3.1 USim Datasets

We evaluate our methods on two USim datasets representing two different text types: ORIGINAL, the USim dataset of Erk et al. (2009), and TWITTER from Gella et al. (2013). Both USim datasets contain pairs of sentences; each sentence in each pair includes a usage of a particular target lemma. Each sentence pair is rated on a scale of 1–5 for how similar in meaning the usages of the target words are in the two sentences.

ORIGINAL consists of sentences from McCarthy and Navigli (2007), which were drawn from a web corpus (Sharoff, 2006). This dataset contains 34 lemmas, including nouns, verbs, adjectives, and adverbs. Each lemma is the target

[2]Inference requires only a single sentence, so the model can infer skip-thought vectors for sentences taken out-of-context, as in the USim datasets.

word in 10 sentences. For each lemma, sentence pairs (SPairs) are formed based on all pairwise comparisons, giving 45 SPairs per lemma. Annotations were provided by three native English speakers, with the average taken as the final gold standard similarity. In a small number of cases the annotators were unable to judge similarity. Erk et al. (2009) removed these SPairs from the dataset, resulting in a total of 1512 SPairs.

TWITTER contains SPairs for ten nouns from ORIGINAL. In this case the "sentences" are in fact tweets. 55 SPairs are provided for each noun. Unlike ORIGINAL, the SPairs are not formed on the basis of all pairwise comparisons amongst a smaller set of sentences. This dataset was annotated via crowd sourcing and carefully cleaned to remove outlier annotations.

3.2 Evaluation

Following Lui et al. (2012) and Gella et al. (2013) we evaluate our systems by calculating Spearman's rank correlation coefficient between the gold standard similarities and the predicted similarities. This enables direct comparison of our results with those reported in these previous studies.

We evaluate our supervised approaches using two cross-validation methodologies. In the first case we apply 10-fold cross-validation, randomly partitioning all SPairs for all lemmas in a given dataset. Using this approach, the test data for a given fold consists of SPairs for target lemmas that were seen in the training data. To determine how well our methods generalize to unseen lemmas, we consider a second cross-validation setup in which we partition the SPairs in a given dataset by lemma. Here the test data for a given fold consists of SPairs for one lemma, and the training data consists of SPairs for all other lemmas.

3.3 Embeddings

We train word2vec's skipgram model on two corpora:[3] (1) a corpus of English tweets collected from the Twitter Streaming APIs[4] from November 2014 to March 2015 containing 1.3 billion tokens; and (2) an English Wikipedia dump from 1 September 2015 containing 2.6 billion tokens. Because of the relatively-low cost of training word2vec, we consider several settings of window size (W=2,5,8) and number of dimensions (D=50,100,300). Embeddings trained on Wikipedia and Twitter are used for experiments on ORIGINAL and TWITTER, respectively.

For the other embeddings we use pre-trained models. We use GloVe vectors from Wikipedia and Twitter, with 300 and 200 dimensions, for experiments on ORIGINAL and TWITTER, respectively.[5] For context2vec we use a 600 dimensional model trained on the ukWaC (Ferraresi et al., 2008), a web corpus of approximately 2 billion tokens.[6] We use a skip-thoughts model with 4800 dimensions, trained on a corpus of books.[7] We use these context2vec and skip-thoughts models for experiments on both ORIGINAL and TWITTER.

D	W	ORIGINAL	TWITTER
50	2	0.251	0.246
50	5	0.262	0.272
50	8	**0.286**	0.282
100	2	0.267	0.248
100	5	0.273	0.253
100	8	0.273	0.298
300	2	0.275	0.266
300	5	0.279	0.295
300	8	0.281	**0.300**

Table 1: Spearman's ρ on each dataset using the unsupervised approach with word2vec embeddings trained using several settings for the number of dimensions (D) and window size (W). The best ρ for each dataset is shown in boldface.

4 Experimental results

We first consider the unsupervised approach using word2vec for a variety of window sizes and number of dimensions. Results are shown in Table 1. All correlations are significant ($p < 0.05$). On both ORIGINAL and TWITTER, for a given number of dimensions, as the window size is increased, ρ increases. Embeddings for larger window sizes tend to better capture semantics, whereas embeddings for smaller window sizes tend to better reflect syntax (Levy and Goldberg, 2014); the

[3] In preliminary experiments the alternative word2vec CBOW model achieved substantially lower correlations than skipgram, and so CBOW was not considered further.

[4] https://dev.twitter.com/

[5] http://nlp.stanford.edu/projects/glove/

[6] https://github.com/orenmel/context2vec

[7] https://github.com/ryankiros/skip-thoughts

Dataset	Embeddings	Unsupervised	Supervised	
			All	Lemma
ORIGINAL	Word2vec	0.281*	0.435*	0.220*
	GloVe	0.218*	0.410*	**0.230***
	Skip-thoughts	0.177*	**0.436***	0.099*
	Context2vec	**0.302***	0.417*	0.172*
TWITTER	Word2vec	**0.300***	**0.384***	**0.196***
	GloVe	0.122*	0.314*	0.134*
	Skip-thoughts	0.095*	0.360*	0.058
	Context2vec	0.122*	0.193*	0.067

Table 2: Spearman's ρ on each dataset using the unsupervised method, and supervised methods with cross-validation folds based on random sampling across all lemmas (All) and holding out individual lemmas (Lemma), for each embedding approach. The best ρ for each experimental setup, on each dataset, is shown in boldface. Significant correlations ($p < 0.05$) are indicated with *.

more-semantic embeddings given by larger window sizes appear to be better-suited to the task of predicting USim. For a given window size, a higher number of dimensions also tends to achieve higher ρ. For example, for a given window size, $D = 300$ gives a higher ρ than $D = 50$ in each case, except for $W = 8$ on ORIGINAL.

The best correlations reported by Lui et al. (2012) on ORIGINAL, and Gella et al. (2013) on TWITTER, were 0.202 and 0.29, respectively. The best parameter settings for our unsupervised approach using word2vec embeddings achieve higher correlations, 0.286 and 0.300, on ORIGINAL and TWITTER, respectively. Lui et al. (2012) and Gella et al. (2013) both report drastic variation in performance for different settings of the number of topics in their models. We also observe some variation with respect to parameter settings; however, any of the parameter settings considered achieves a higher correlation than Lui et al. (2012) on ORIGINAL. For TWITTER, parameter settings with $W \geq 5$ and $D \geq 100$ achieve a correlation comparable to, or greater than, the best reported by Gella et al. (2013)

We now consider the unsupervised approach, using the other embeddings. Based on the previous findings for word2vec, we only consider this model with $W = 8$ and $D = 300$ here. Results are shown in Table 2 in the column labeled "Unsupervised". For ORIGINAL, context2vec performs best (and indeed outperforms word2vec for all parameter settings considered). This result demonstrates that approaches to predicting USim that explicitly embed the context of a target word can outperform approaches based on averaging word embeddings (i.e., word2vec and GloVe) or embedding sentences (skip-thoughts). This result is particularly strong because we consider a range of parameter settings for word2vec, but only used the default settings for context2vec.[8] Word2vec does however perform best on TWITTER. The relatively poor performance of context2vec and skip-thoughts here could be due to differences between the text types these embedding models were trained on and the evaluation data. GloVe performs poorly, even though it was trained on tweets for these experiments, but that it performs less well than word2vec is consistent with the findings for ORIGINAL.

Turning to the supervised approach, we first consider results for cross-validation based on randomly partitioning all SPairs in a dataset (column "All" in Table 2). The best correlation on TWITTER (0.384) is again achieved using word2vec, while the best correlation on ORIGINAL (0.434) is obtained with skip-thoughts. The difference in performance amongst the various embedding approaches is, however, somewhat less here than in the unsupervised setting. For each embedding approach, and each dataset, the correlation in the supervised setting is better than that in the unsupervised setting, suggesting that if labeled training data is available, supervised approaches can give substantial improvements over unsupervised approaches to predicting USim.[9] However, this experimental setup does not show the extent to which the supervised approach is able to generalize to previously-unseen lemmas.

The column labeled "Lemma" in Table 2 shows results for the supervised approach for cross-validation using lemma-based partitioning. In these experiments, the test data consists of usages of a target lemma that was not seen as a target lemma during training. For each dataset, the correlations achieved here for each type of embedding are lower than those of the corresponding unsupervised method, with the exception of GloVe. In

[8]The context2vec model has 600 dimensions, and was trained on the ukWac, whereas our word2vec model for ORIGINAL is trained on Wikipedia. To further compare these approaches we also trained word2vec on the ukWaC with 600 dimensions and a window size of 8. These word2vec settings also did not outperform context2vec.

[9]These results on ORIGINAL must be interpreted cautiously, however. The same sentences, albeit in different SPairs, occur in both the training and testing data for a given fold. This issue does not affect TWITTER.

the case of ORIGINAL, the higher correlation for GloVe relative to the unsupervised setup appears to be largely due to improved performance on adverbs. Nevertheless, for each dataset, the correlations achieved by GloVe are still lower than those of the best unsupervised method on that dataset. These results demonstrate that the supervised approach generalizes poorly to new lemmas. This negative result indicates an important direction for future work — identifying strategies to training supervised approaches to predicting USim that generalize to unseen lemmas.

5 Conclusions

Word senses are not discrete, and multiple senses are often applicable for a given usage of a word. Moreover, for text types that have a relatively-high rate of out-of-vocabulary words, such as social media text, many words will be missing from sense inventories. USim is an approach to determining word meaning in context that does not rely on a sense inventory, addressing these concerns.

We proposed unsupervised approaches to USim based on embeddings for words, contexts, and sentences. We achieved state-of-the-art results over USim datasets based on Twitter text and more-conventional texts. We further considered supervised approaches to USim based on these same methods for forming embeddings, and found that although these methods outperformed the unsupervised approaches, they performed poorly on lemmas that were unseen in the training data.

The approaches to learning word embeddings that we considered (word2vec and GloVe) both learn a single vector representing each word type. There are, however, approaches that learn multiple embeddings for each type that have been applied to predict word similarity in context (Huang et al., 2012; Neelakantan et al., 2014, for example). In future work, we intend to also evaluate such approaches for the task of predicting usage similarity. We also intend to consider alternative strategies to training supervised approaches to USim in an effort to achieve better performance on unseen lemmas.

Acknowledgments

This research was financially supported by the Natural Sciences and Engineering Research Council of Canada, the New Brunswick Innovation Foundation, ACENET, and the University of New Brunswick.

References

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources? In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, pages 356–364, Nagoya, Japan.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2009. Investigations on word senses and word usages. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 10–18, Singapore.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop: Can we beat Google*, pages 47–54, Marrakech, Morocco.

Spandana Gella, Paul Cook, and Bo Han. 2013. Unsupervised word usage similarity in social media texts. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 248–253, Atlanta, USA.

Eric Huang, Richard Socher, Christopher Manning, and Andrew 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 (Volume 1: Long Papers)*, pages 873–882, Jeju Island, Korea.

Samuel Johnson. 1755. *A Dictionary of the English Language*. London.

David Jurgens and Ioannis Klapaftis. 2013. Semeval-2013 task 13: Word sense induction for graded and non-graded senses. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, pages 290–299, Atlanta, USA.

Adam Kilgarriff. 1997. "I Don't Believe in Word Senses". *Computers and the Humanities*, 31(2):91–113.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. In C. Cortes, N. D. Lawrence, D. D. Lee,

M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3276–3284. Curran Associates, Inc.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Comouational Linguistics*, pages 302–308, Maryland, USA.

Marco Lui, Timothy Baldwin, and Diana McCarthy. 2012. Unsupervised estimation of word usage similarity. In *Proceedings of the Australasian Language Technology Association Workshop 2012*, pages 33–41, Dunedin, New Zealand.

Diana McCarthy and Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional lstm. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at the International Conference on Learning Representations, 2013*, Scottsdale, USA.

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 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1059–1069, Doha, Qatar.

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

Serge Sharoff. 2006. Open-source corpora: Using the net to fish for linguistic data. *Corpus Linguistics*, 11(4):435–462.