A Word Embedding Approach to Predicting the Compositionality of Multiword Expressions

Bahar Salehi,♠ Paul Cook♡ and Timothy Baldwin♠
♠ NICTA Victoria Research Laboratory
Department of Computing and Information Systems
The University of Melbourne
Victoria 3010, Australia

♡ Faculty of Computer Science
University of New Brunswick
Fredericton, NB E3B 5A3, Canada

bsalehi@student.unimelb.edu.au, paul.cook@unb.ca, tb@ldwin.net

Abstract

This paper presents the first attempt to use word embeddings to predict the compositionality of multiword expressions. We consider both single- and multi-prototype word embeddings. Experimental results show that, in combination with a back-off method based on string similarity, word embeddings outperform a method using count-based distributional similarity. Our best results are competitive with, or superior to, state-of-the-art methods over three standard compositionality datasets, which include two types of multiword expressions and two languages.

1 Introduction

Multiword expressions (MWEs) are word combinations that display some form of idiomaticity (Baldwin and Kim, 2009), including semantic idiomaticity, wherein the semantics of the MWE (e.g. *ivory tower*) cannot be predicted from the semantics of the component words (e.g. *ivory* and *tower*). Recent NLP work on semantic idiomaticity has focused on the task of "compositionality prediction", in the form of a regression task whereby a given MWE is mapped onto a continuous-valued compositionality score, either for the MWE as a whole or for each of its component words (Reddy et al., 2011; Schulte im Walde et al., 2013; Salehi et al., 2014b).

Separately in NLP, there has been a recent surge of interest in learning distributed representations of word meaning, in the form of "word embeddings" (Collobert and Weston, 2008; Mikolov et al., 2013a) and composition over distributed representations (Socher et al., 2012; Baroni et al., 2014).

This paper is the first attempt to bring together the work on word embedding-style distributional analysis with compositionality prediction of MWEs. In the context of compositionality prediction, our primary research questions here are:

RQ1: Are word embeddings superior to conventional count-based models of distributional similarity?
RQ2: How sensitive to parameter optimisation are different word embedding approaches?
RQ3: Are multi-prototype word embeddings empirically superior to single-prototype word embeddings?

We explore these questions relative to three compositionality prediction datasets spanning two MWE construction types (noun compounds and verb particle constructions) and two languages (English and German), and arrive at the following conclusions: (1) consistent with recent work over other NLP tasks, word embeddings are superior to count–based models of distributional similarity (and also translation-based string similarity); (2) the results are relatively stable under parameter optimisation for a given word embedding learning approach; and (3) based on two simple approaches to composition, single word embeddings are empirically slightly superior to multi-prototype word embeddings overall.

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 977–983,
Denver, Colorado, May 31 – June 5, 2015.

2 Related Work

Recent work on distributed approaches to distributional semantics has demonstrated their utility in a wide range of NLP tasks, including identifying various morphosyntactic and semantic relations (Mikolov et al., 2013a), dependency parsing (Bansal et al., 2014), sentiment analysis (Socher et al., 2013), named-entity recognition (Collobert and Weston, 2008; Passos et al., 2014), and machine translation (Zou et al., 2013; Devlin et al., 2014). Despite the wealth of research applying word embeddings within NLP, they have not yet been considered for predicting the compositionality of MWEs.

Much prior work on MWEs has been tailored to specific kinds of MWEs in particular languages (e.g. English verb–noun combinations (Fazly et al., 2009)). There has however been recent interest in approaches to MWEs that are more broadly applicable to a wider range of languages and MWE types (Brooke et al., 2014; Salehi et al., 2014b; Schneider et al., 2014). Word embeddings could form the basis for such an approach to predicting MWE compositionality.

3 Methodology

In this work, we estimate the compositionality of an MWE based on the similarity between the expression and its component words in vector space. We use three different vector-space models: (1) a simple count-based model of distributional similarity; (2) word embeddings based on WORD2VEC; and (3) a multi-sense skip-gram model that, unlike the previous two models, is able to learn multiple embeddings per target word (or MWE). For all three models, we first greedily pre-tokenise the corpus to represent each MWE as a single token, similarly to Baldwin et al. (2003). In this, we apply the constraint that no language-specific pre-processing can be applied to the training corpus, in order to make the method maximally language independent. As such, we cannot perform any form of lemmatisation, and MWE identification takes the form of simple string match for concatenated instances of the component words, naively assuming that all occurrences of that word combination are MWEs. We detail each of the distributional similarity methods below.

3.1 Count-Based Distributional Similarity

Our first method for building vectors is that of Salehi et al. (2014b): the top 50 most-frequent words in the training corpus are considered to be stopwords and discarded, and words with frequency rank 51–1051 are considered to be the content-bearing words, which form the dimensions for our vectors, in the manner of Schütze (1997). To measure the similarity of the MWE vector and the component word vectors, we considered two different approaches.

The first approach is based on Reddy et al. (2011) and Schulte im Walde et al. (2013). The similarity between the MWE and each of its components is measured, and the overall compositionality of the MWE is computed by combining the similarity scores for the two components as follows:

$$comp_1(\mathbf{MWE}) = \alpha sim(\mathbf{MWE}, \mathbf{C}_1) + (1-\alpha) sim(\mathbf{MWE}, \mathbf{C}_2)$$

where $\mathbf{MWE}$ is the vector associated with the MWE, $\mathbf{C}_i$ is the vector associated with the ith component word of the MWE, sim is a vector similarity function, and $\alpha \in [0, 1]$ is a weight parameter.

We also experimented with the approach from Mitchell and Lapata (2010), where $\mathbf{MWE}$ is compared directly with a composed vector of the component words, based on vector addition:[1]

$$comp_2(\mathbf{MWE}) = sim(\mathbf{MWE}, \mathbf{C}_1 + \mathbf{C}_2)$$

For both $comp_1$ and $comp_2$, we used cosine similarity as our similarity measure sim.

3.2 WORD2VEC

Our second method is based on the recurrent neural network language model (RNNLM) approach to learning word embeddings of Mikolov et al. (2013a) and Mikolov et al. (2013b), using the WORD2VEC package.[2] WORD2VEC uses a log-linear model inspired by the original RNNLM approach of Mikolov et al. (2010), in two forms: (1) a continuous bag-of-words ("CBOW") model, whereby all words in a context window are averaged in a single projection layer; and (2) a continuous skip-gram model

[1] We also experimented with vector multiplication, but found it to perform poorly compared to the other approaches.

[2] https://code.google.com/p/word2vec/

("C-SKIP"), whereby a given word in context is projected onto a projection layer, and used to predict its immediate context (preceding and following words). WORD2VEC generates a vector of fixed dimensionality d for each pre-tokenised word/MWE type with frequency above a certain threshold in the training corpus. We again use $comp_1$ and $comp_2$ to estimate compositionality from these vectors.

3.3 Multi-Sense Skip-gram Model

One potential shortcoming of WORD2VEC is that it generates a single word embedding for each word, irrespective of the relative polysemy of the word. Neelakantan et al. (2014) proposed a method motivated by WORD2VEC, which efficiently learns multiple embeddings per word/MWE. We refer to this approach as the multi-sense skip-gram (MSSG) model. We once again compose the resultant vectors with $comp_1$ and $comp_2$, but modify the formulation slightly to handle the variable number of vectors for each word/MWE, by searching over the cross-product of vectors in each sim calculation and taking the maximum in each case. We initially set the number of embeddings to 2 in our MSSG experiments — in keeping with the findings in Neelakantan et al. (2014) — but come back to examine the impact of the number of embeddings on compositionality prediction in Section 5.

4 Datasets

We evaluate our methods over three datasets:[3] (1) English noun compounds ("ENCs", e.g. *spelling bee* and *swimming pool*); (2) English verb particle constructions ("EVPCs", e.g. *stand up* and *give away*); and (3) German noun compounds ("GNCs", e.g. *ahornblatt* "maple leaf" and *eidechse* "lizard").

The ENC dataset consists of 90 binary English noun compounds, and is annotated on a continuous $[0, 5]$ scale for both overall compositionality and the component-wise compositionality of each of the modifier and head noun (Reddy et al., 2011). The state-of-the-art method for this dataset (Salehi et al., 2014b) is a supervised support vector regression model, trained over the distributional method from Section 3.1 as applied to both English and 51 target languages (under word and MWE translation).

The EVPC dataset consists of 160 English verb particle constructions, and is manually annotated for compositionality on a binary scale for each of the head verb and particle (Bannard, 2006). In order to translate the dataset into a regression task, we calculate the overall compositionality as the number of annotations of entailment for the verb, divided by the total number of verb annotations for that VPC. The state-of-the-art method for this dataset (Salehi et al., 2014b) is a linear combination of: (1) the distributional method from Section 3.1; (2) the same method applied to 10 target languages (under word and MWE translation, selecting the languages using supervised learning); and (3) the string similarity method of Salehi and Cook (2013).

The GNC dataset consists of 246 German noun compounds, and is annotated on a continuous $[1, 7]$ scale (von der Heide and Borgwaldt, 2009; Schulte im Walde et al., 2013). The state-of-the-art method for this dataset is a distributional similarity method applied to part-of-speech tagged and lemmatised data (Schulte im Walde et al., 2013).

5 Experiments

For all experiments, we train our models over raw text Wikipedia corpora for either English or German, depending on the language of the dataset. The raw English and German corpora were preprocessed using the WP2TXT toolbox[4] to eliminate XML and HTML tags and hyperlinks, and punctuation was removed. Finally, word-tokenisation was performed based on simple whitespace delimitation, after which we greedily identified all string occurrences of the MWEs in each of our datasets and combined them into a single token.[5]

The word embedding approaches are unable to generate vector representations for tokens which occur with frequency below a fixed cutoff.[6] In order to

[3] We also considered using the dataset from the DisCo shared task (Biemann and Giesbrecht, 2011), but ultimately excluded it because it includes different types of MWEs without indication of the syntactic type of a given MWE, preventing us from carrying out construction-specific parameter tuning.

[4] http://wp2txt.rubyforge.org/

[5] For English, a single model was trained over a corpus containing both ENC and EVPC tokens.

[6] For a frequency threshold of 15, the total numbers of ENCs, EVPCs and GNCs for which we were unable to generate word embeddings were 3, 0 and 25, respectively, in the latter case, largely as a result of our simple tokenisation strategy and

Dataset	Method		$comp_1$	$comp_1$ +SS	$comp_2$	$comp_2$ +SS
ENC	WORD2VEC	($d = 500$, C-SKIP)	.628	.761	.632	.761
		($d = 500$, CBOW)	.696	.786	.710	.791
		($d = 1000$, C-SKIP)	.636	.764	.648	.767
		($d = 1000$, CBOW)	.717	.789	.736	**.796**
	MSSG	($d = 300, w = 5$)	.640	.764	.624	.759
		($d = 600, w = 5$)	.615	.758	.594	.758
		($d = 600, w = 10$)	.614	.749	.631	.756
	Distributional similarity			.714		
	String similarity			.644		
	State-of-the-art			.744		
EVPC	WORD2VEC	($d = 500$, C-SKIP)	.289	.496	—	—
		($d = 500$, CBOW)	.293	.486	—	—
		($d = 1000$, C-SKIP)	.289	.504	—	—
		($d = 1000$, CBOW)	.289	.489	—	—
	MSSG	($d = 300, w = 5$)	.309	**.506**	—	—
		($d = 600, w = 5$)	.294	.498	—	—
		($d = 600, w = 10$)	.273	.494	—	—
	Distributional similarity			.165		
	String similarity			.385		
	State-of-the-art			.417		
GNC	WORD2VEC	($d = 500$, C-SKIP)	.393	.442	.321	.415
		($d = 500$, CBOW)	.400	.439	.361	.423
		($d = 1000$, C-SKIP)	.341	.411	.282	.394
		($d = 1000$, CBOW)	.371	.414	.349	.411
	MSSG	($d = 300, w = 5$)	.181	.320	.122	.295
		($d = 600, w = 5$)	.202	.335	.146	.303
		($d = 600, w = 10$)	.155	.310	.101	.282
	Distributional Similarity			.140		
	String Similarity			.372		
	State-of-the-art			**.450**		

Table 1: Pearson's correlation (r) for the different methods over the three datasets; the state-of-the-art for each dataset is described in Section 4

generate a compositionality prediction back-off for the small numbers of MWEs in this category, we assign a default value, which is the mean of computed compositionality scores for other instances.[7]

As a baseline, we use the translation string similarity approach of Salehi and Cook (2013), including the cross-validation-based method for selecting the 10 best languages to use for each dataset. We further include a linear combination of the string similarity method with each of the various approaches based on word embeddings.

Table 1 shows the results for the various methods, over a range of hyper-parameter settings for each of WORD2VEC (vector dimensionality d; we also present results for CBOW vs. C-SKIP) and MSSG (vector dimensionality d and window size w), informed by the experimental results in the respective publications. Note that for EVPC, we don't use the vector for the particle, in keeping with Salehi et al. (2014b); as such, there are no results for $comp_2$. For $comp_1$, α is set to 1.0 for EVPC, and 0.7 for both ENC and GNC, also based on the findings of Salehi et al. (2014b).

The results indicate that the approaches using both WORD2VEC and MSSG outperform simple distributional and string similarity by a substantial margin. Further, over a variety of parameteriza-

lack of lemmatisation.

[7]We also experimented with using the string similarity approach as a back-off, which resulted in marginally lower results than what is reported in Table 1.

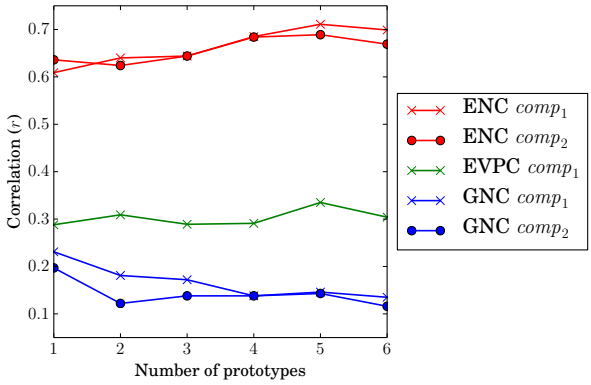

Figure 1: The effect of the number of prototypes on the results with MSSG

tions, they surpass the state-of-the-art methods for ENC and EVPC; in the case of GNC, the best-performing method (WORD2VEC with $d = 500$ and C-SKIP) roughly matches the state-of-the-art. Note that in each case, the state-of-the-art is achieved using varying levels of supervision over labelled data (ENC and EVPC) or language-specific pre-processing (GNC), whereas the word embedding methods use no labelled data. As such, the answer to RQ1 would appear to be a resounding yes.

Looking to RQ2, the models are remarkably insensitive to hyper-parameter optimisation for EVPC, but there are slight deviations in the results for ENC and GNC. Having said that, they are largely between the different word embedding approaches, and the results for a given approach under different parameter settings is relatively stable. A large part of the cause of the drop in results and greater parameter sensitivity over GNC is the lower token frequencies, through a combination of the Wikipedia corpus being markedly smaller and our naive tokenisation strategy having low recall over German due to the richer morphology. As such, the answer would appear to be a tentative "relatively insensitive, assuming high token frequencies".

Finally, looking to RQ3, there was little separating WORD2VEC and MSSG over ENC, but over the other two datasets, WORD2VEC had a clear advantage. Given the high levels of polysemy observed in high frequency English verb particle constructions (Salehi et al., 2014a), this result for EVPC was particularly surprising, and suggests that, at least under our two basic forms of composition, multi-prototype word embeddings are at best equal to, and in many cases, inferior to, single-prototype word embeddings.

According to the results, the string similarity approach complements all word-embedding approaches. We hypothesise that this is because it is not based on any corpus, and is thus not biased by the frequency of token instances in the corpus.

In Table 1, the number of embeddings for MSSG was set to 2 prototypes, based on the default recommendations of Neelakantan et al. (2014). To investigate the impact of this parameter on our results, we retrained MSSG over the range $[1, 6]$ and reran our experiments for each set of embeddings over the three datasets (without string similarity, to isolate the effect of the number of embeddings), as shown in Figure 1. For both English datasets (ENC and EVPC), setting the number of prototypes to a value higher than 2 boosts the results slightly, with 5 prototypes appearing to be the optimal value. For the German dataset (GNC), on the other hand, the best results are actually achieved for a single prototype. Further research is required to better understand this effect.

6 Conclusions

We presented the first approach to using word embeddings to predict the compositionality of MWEs. We showed that this approach, in combination with information from string similarity, surpassed, or was competitive with, the current state-of-the-art on three compositionality datasets. In future work we intend to explore the contribution of information from word embeddings of a target expression and its component words under translation into many languages, along the lines of Salehi et al. (2014b).

Acknowledgements

We thank the anonymous reviewers for their insightful comments and valuable suggestions. NICTA is funded by the Australian government as represented by Department of Broadband, Communication and Digital Economy, and the Australian Research Council through the ICT Centre of Excel-

lence programme.

References

Timothy Baldwin and Su Nam Kim. 2009. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*. CRC Press, Boca Raton, USA, 2nd edition.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96, Sapporo, Japan.

Colin James Bannard. 2006. *Acquiring Phrasal Lexicons from Corpora*. Ph.D. thesis, University of Edinburgh.

Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. Tailoring continuous word representations for dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 809–815, Baltimore, USA.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pages 238–247, Baltimore, USA.

Chris Biemann and Eugenie Giesbrecht. 2011. Distributional semantics and compositionality 2011: Shared task description and results. In *Proceedings of the workshop on distributional semantics and compositionality*, pages 21–28, Portland, Oregon, USA.

Julian Brooke, Vivian Tsang, Graeme Hirst, and Fraser Shein. 2014. Unsupervised multiword segmentation of large corpora using prediction-driven decomposition of n-grams. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 753–761, Dublin, Ireland.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning (ICML 2008)*, pages 160–167, Helsinki, Finland.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pages 1370–1380, Baltimore, USA.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*, pages 1045–1048, Makuhari, Japan.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

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 2014)*, pages 1059–1069, Doha, Qatar.

Alexandre Passos, Vineet Kumar, and Andrew McCallum. 2014. Lexicon infused phrase embeddings for named entity resolution. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 78–86, Baltimore, USA.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of IJCNLP*, pages 210–218, Chiang Mai, Thailand.

Bahar Salehi and Paul Cook. 2013. Predicting the compositionality of multiword expressions using translations in multiple languages. 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 266–275, Atlanta, Georgia, USA, June.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2014a. Detecting non-compositional MWE components using Wiktionary. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1792–1797, Doha, Qatar, October.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2014b. Using distributional similarity of multi-way transla-

tions to predict multiword expression compositionality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 472–481, Gothenburg, Sweden, April.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. 2014. Discriminative lexical semantic segmentation with gaps: Running the mwe gamut. *Transactions of the Association of Computational Linguistics*, 2:193–206.

Sabine Schulte im Walde, Stefan Müller, and Stephen Roller. 2013. Exploring vector space models to predict the compositionality of German noun-noun compounds. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, pages 255–265, Atlanta, USA.

Hinrich Schütze. 1997. *Ambiguity Resolution in Language Learning*. CSLI Publications, Stanford, USA.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning 2012 (EMNLP-CoNLL 2012)*, pages 1201–1211, Jeju Island, Korea.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, pages 1631–1642, Seattle, USA.

Claudia von der Heide and Susanne Borgwaldt. 2009. Assoziationen zu Unter, Basis und Oberbegriffen. Eine explorative Studie. In *Proceedings of the 9th Norddeutsches Linguistisches Kolloquium*, pages 51–74.

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398, Seattle, USA.