When a Red Herring is Not a Red Herring: Using Compositional Methods to Detect Non-Compositional Phrases

Julie Weeds, Thomas Kober, Jeremy Reffin and **David Weir**
TAG Laboratory, Department of Informatics
University of Sussex, Brighton, BN1 9QH, UK
{J.E.Weeds, T.Kober, J.P.Reffin, D.J.Weir}@sussex.ac.uk

Abstract

Non-compositional phrases such as *red herring* and weakly compositional phrases such as *spelling bee* are an integral part of natural language (Sag et al., 2002). They are also the phrases that are difficult, or even impossible, for good compositional distributional models of semantics. Compositionality detection therefore provides a good testbed for compositional methods. We compare an integrated compositional distributional approach, using sparse high dimensional representations, with the adhoc compositional approach of applying simple composition operations to state-of-the-art neural embeddings.

1 Introduction

One current focus within the field of distributional semantics is enabling systems to make inferences about phrase-level or sentence-level similarity. One popular approach (Mitchell and Lapata, 2010) is to build phrase or sentence-level representations by composing word-level representations and then measuring similarity directly. Success is usually measured in terms of correlation with human similarity judgments. However, evaluating measures of phrase-level similarity directly against human judgments of similarity ignores the problem that it is not always possible to determine meaning in a compositional manner. If we compose the meaning representations for *red* and *herring*, we might expect to get a very different representation from the one which could be directly inferred from corpus observations of the phrase *red herring*. Thus any judgements of the similarity of two composed phrases may be confounded by the degree to which those phrases are compositional.

In this paper, we use a compound noun compositionality dataset (Reddy et al., 2011) to investigate the extent to which the underlying definition of context has an effect on a model's ability to support composition. We compare the Anchored Packed Tree (APT) model (Weir et al., 2016), where composition is an integral part of the distributional model, with the commonly employed approach of applying naïve compositional operations to state-of-the-art distributional representations.

2 Background

Context definition	Example features
Proximity (+-2)	*recently*, *graduated, folded*
Typed dep. rel.	$\langle$NMOD, $graduated\rangle$, $\langle\overline{\text{NSUBJ}}$, $folded\rangle$
Untyped dep. rel.	*graduated, folded*
Typed dep. path	$\langle$NMOD, $graduated\rangle$, $\langle\overline{\text{NSUBJ}}$, $folded\rangle$, $\langle\overline{\text{NSUBJ}}$.DOBJ, $clothes\rangle$, $\langle$NMOD.AMOD, $recently\rangle$, $\langle\overline{\text{NSUBJ}}$.DOBJ.AMOD, $dry\rangle$
Untyped dep. path	*recently*, *graduated*, *folded*, *dry*, *clothes*

Table 1: Possible contextual features of *student*

Consider the occurrence of the word *student* in the sentence *"The recently graduated student folded the dry clothes."* Different distributional representations leverage the context, e.g., the fact that the target word *student* has occurred in the context *folded*, in different ways. Table 1 illustrates the contextual features which might be generated for *student* given different definitions of context. The most commonly used definition of context, in both traditional count-based representations and in more recent distributed embeddings, is proximity, i.e., the contextual features of a word occurrence are all those words which occur within a certain context window around the occurrence. However, contextual features may also be defined

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 529–534,
Valencia, Spain, April 3-7, 2017. ©2017 Association for Computational Linguistics

in terms of dependency relations. For example, in a dependency parse of the sentence we would expect to see a direct-object relation from *folded* to *student*. Contextual features based on dependency relations may be typed (i.e., include the name of the dependency relation) or untyped (Baroni and Lenci, 2010). Padó and Lapata (2007) proposed using dependency paths to define untyped contextual features; here any word in the context which has a dependency path to the target is considered a contextual feature. Weeds et al. (2014) proposed using dependency paths to define typed contextual features which could be used to align representations before composition. This idea is further refined in the APT framework of Weir et al. (2016).

Naïve composition of distributional representations, e.g., using pointwise addition and multiplication, has proved very popular and effective. In an evaluation across 3 different benchmark tasks (Dinu et al., 2013), the lexical function model (Baroni and Zamparelli, 2010) was shown to be consistently the best-performing, but in the composition of adjective-noun phrases, simple additive and multiplicative models were highly competitive. Milajevs et al. (2014) compared neural word representations with count-based vectors on 4 different tasks using a variety of naïve and tensor-based compositional models. The neural word representations consistently outperformed the traditional count-based vectors. Considering the results for the neural word representations, pointwise addition outperformed all of the other compositional models considered on 3 of the tasks.

Typed distributional representations cannot be straightforwardly composed using naïve operations (Weeds et al., 2014). The APT approach (Weir et al., 2016) overcomes this problem by defining contextual features in terms of complete dependency paths and then ensuring that the representations of target words are properly aligned before composition. For example, to carry out the composition of *student* with *folded* in the example sentence, it is necessary to align the representations. This can be done by offsetting all of the features of *student* by its dependency relation (NSUBJ) with *folded*. Intuitively we are viewing the representation of *student* from the perspective of actions (i.e., verbs) which are likely to be carried out by students. This view can be straightforwardly composed with the representation of *folded* because the representations are aligned i.e., they have features of the same type (e.g., DOBJ).

3 Compositionality of compound nouns

Compositionality detection (Reddy et al., 2011) involves deciding whether a given multiword expression is compositional or not i.e., whether the meaning can be understood from the literal meaning of its parts. Reddy et al. (2011) introduced a dataset consisting of 90 compound nouns along with human judgments of their literality or compositionally at both the constituent and the phrase level. All judgments are given on a scale of 0 to 5, where 5 is high. For example, the phrase *spelling bee* is deemed to have high literalness in its use of the first constituent, low literalness in its use of the second constituent and a medium level of literalness with respect to the whole phrase.

Assuming the distributional hypothesis (Harris, 1954), the observed co-occurrences of compositional target phrases are highly likely to have occurred with one or both of the constituents independently. On the other hand, the observed co-occurrences of non-compositional target phrases are much less likely to have occurred with either of the constituents independently. Thus, a good compositionality function, without any access to the observed co-occurrences of the target phrases, is highly likely to return vectors which are similar to observed phrasal vectors for compositional phrases but much less likely to return similar vectors for non-compositional phrases. Accordingly, as observed elsewhere (Reddy et al., 2011; Salehi et al., 2015; Yazdani et al., 2015), compositional methods can be evaluated by correlating the similarity of composed and observed phrase representations with the human judgments of compositionality. A similar idea is also explored by Kiela and Clark (2013) who detect non-compositional phrases by comparing the neighbourhoods of phrases where individual words have been substituted for similar words.

Reddy et al. (2011) carried out experiments with a vector space model built from ukWaC (Ferraresi et al., 2008) using untyped co-occurrences (window size=100). Used 3-fold cross-validation, they found that using weighted addition outperformed multiplication as a compositionality function. With their optimal settings, they achieved a Spearman's rank correlation coefficient of 0.714 with the human judgments, which remains the state-of-the-art on this dataset[1]. For consistency with the experiments of Reddy et al. (2011), the corpus used in this experiment is the same fully-annotated version of the web-derived ukWaC corpus (Ferraresi et al., 2008). This corpus has been tokenised, POS-tagged and lemmatised with TreeTagger (Schmid, 1994) and dependency-parsed with the Malt Parser (Nivre, 2004). It contains about 1.9 billion tokens.

In order to create a corpus which contains compound nouns, we further preprocessed the corpus by identifying occurrences of the 90 target compound nouns and recombining them into a single lexical item. We then created a number of elementary representations for every token in the corpus.

3.1 Untyped contextual features

For each word and compound phrase, neural representations were constructed using the word2vec tool (Mikolov et al., 2013). Whilst it is not possible or appropriate to carry out an exhaustive parameter search, we experiment with a number of commonly used and recommended parameter settings. We investigate both the `cbow` and `skip-gram` models with 50, 100 and 300 dimensions and experiment with the subsampling threshold, trying 10^{-3}, 10^{-4} and 10^{-5}. As recommended in the documentation, we use a window size of 5 for `cbow` and of 10 for `skip-gram`. Early experiments with different composition operations, showed `add` to be the only promising option. Similarity between composed and observed representations is computed using the cosine measure.

3.2 Typed contextual features

For each word and compound phrase, elementary APT representations were constructed using the method and recommended settings of Weir et al. (2016). For efficiency, we did not consider paths of length 3 or more. In relation to the construction of the elementary APTs, the most obvious parameter is the nature of the weight associated with each feature. We consider both the use of probabilities[2] and positive pointwise mutual information (PPMI) values. Levy et al. (2015) showed that the use of context distribution smoothing ($\alpha = 0.75$) in the PMI calculation can lead to performance comparable with state-of-the-art word embeddings on word similarity tasks. We use this modified definition of PMI and experiment with $\alpha = 0.75$ and $\alpha = 1$.[3]

Having constructed elementary APTs, the APT composition process involves aligning and composing these elementary APTs. We investigate using $\bigsqcup_{\text{INT}}$, which takes the minimum of each of the constituent's feature values and $\bigsqcup_{\text{UNI}}$, which performs pointwise addition. Following Reddy et al. (2011), when using the $\bigsqcup_{\text{UNI}}$ operation, we experiment with weighting the contributions of each constituent to the composed APT representation using the parameter, h. For example, if $\mathbf{A}_2$ is the APT associated with the head of the phrase and $\mathbf{A}_1^{\delta}$ is the properly aligned APT associated with the modifier where δ is the dependency path from the head to the modifier (e.g. NMOD or AMOD), the composition operations can be defined as:

$$\bigsqcup_{\text{INT}} \left\{ \mathbf{A}_1^{\delta}, \mathbf{A}_2 \right\} \tag{1}$$

$$\bigsqcup_{\text{UNI}} \left\{ (1-h)\mathbf{A}_1^{\delta}, h\mathbf{A}_2 \right\} \tag{2}$$

We have also considered composition without alignment of the modifier's APT, i.e, using $\mathbf{A}_1$:

$$\bigsqcup_{\text{INT}} \{ \mathbf{A}_1, \mathbf{A}_2 \} \tag{3}$$

$$\bigsqcup_{\text{UNI}} \{ (1-h)\mathbf{A}_1, h\mathbf{A}_2 \} \tag{4}$$

In general, one would expect there to be little overlap between APTs which have not been properly aligned. However, in the case where δ is the NMOD relation, i.e., the internal relation in the vast majority of the compound phrases, both modifier and head are nouns and therefore there may well be considerable overlap between their unaligned dependency features. In order to examine the contribution of both the aligned and unaligned APTs in the composition process, we used a hybrid method where the composed representation is defined as:

$$\bigsqcup_{\text{INT}} \left\{ (q\mathbf{A}_1^{\delta} + (1-q)\mathbf{A}_1), \mathbf{A}_2 \right\} \tag{5}$$

[1] Hermann et al. (2012) proposed using generative models for modeling the compositionality of noun-noun compounds. Using interpolation to mitigate the sparse data problem, their model beat the baseline of weighted addition on the Reddy et al. (2011) evaluation task when trained on the BNC. However, these results were still significantly lower than those reported by Reddy et al. (2011) using the larger ukWaC corpus.

[2] referred to as normalised counts by Weir et al. (2016)

[3] $\alpha = 1$ corresponds to the standard definition of PMI used elsewhere.

Embedding method	$t = 10^{-3}$	$t = 10^{-4}$	$t = 10^{-5}$
`cbow`, 50d	0.73	0.65	0.62
`cbow`, 100d	**0.74**	0.65	0.64
`cbow`, 300d	0.70	0.70	0.67
`skip-gram`, 50d	0.59	0.64	0.62
`skip-gram`, 100d	0.62	0.64	0.64
`skip-gram`, 300d	0.63	0.64	**0.68**

Table 2: Average ρ using neural word embeddings

Compositional Model	PPMI $\alpha = 1$		PPMI $\alpha = 0.75$	
	`CF`	`CS`	`CF`	`CS`
Aligned $\bigsqcup_{\text{INT}}$ (Eq. 1)	0.72	0.70	**0.75**	0.72
Aligned $\bigsqcup_{\text{UNI}}$ (Eq. 2)	0.71	0.72	0.72	**0.75**
Unaligned $\bigsqcup_{\text{INT}}$ (Eq. 3)	0.74	0.72	0.72	0.73
Unaligned $\bigsqcup_{\text{UNI}}$ (Eq. 4)	0.77	0.75	**0.78**	0.77
Hybrid $\bigsqcup_{\text{INT}}$ (Eq. 5)	0.74	0.73	0.73	0.73
Hybrid $\bigsqcup_{\text{UNI}}$ (Eq. 6)	0.78	0.78	**0.79**	0.76

Table 3: Average ρ using APT representations.

$$\bigsqcup_{\text{UNI}} \left\{ (1-h)(q\mathbf{A}_1^{\delta} + (1-q)\mathbf{A}_1), h\mathbf{A}_2 \right\} \qquad (6)$$

In the case where representations consist of APT weights which are probabilities, PPMI is estimated after composition. Therefore we refer to this as compose-first (`CF`) in contrast to compose-second (`CS`) where composition is carried out after PPMI calculations. In both cases, the cosine measure is applied to vectors made up PPMI values in order to calculate the similarity of the observed and composed representations.

4 Results

We used repeated 3-fold cross-validation to enable us to estimate[4] the model parameters h and q. Results for all models are then reported in terms of average Spearman rank correlation scores (ρ) of phrase compositionality scores with human judgements on the corresponding testing samples. We used a sufficiently large number of repetitions that errors are all small (≤ 0.0015) and thus any difference observed which is greater than 0.005 is statistically significant at the 95% level. Boldface is used to indicate the best performing configuration of parameters for a particular model.

Table 2 summarises results for different parameter settings for the neural word embeddings. Looking at the results in Table 2, we see that the `cbow` model significantly outperforms the `skip-gram` model. Using the `cbow` model with 100 dimensions and a subsampling threshold of $t = 10^{-3}$ gives a performance of 0.74 which is significantly higher than the previous state-of-the-art reported in Reddy et al. (2011). Since both of these models are based on untyped co-occurrences, this performance gain can be seen as the result of implicit parameter optimisation.

Table 3 summarises results for different composition operations and parameter settings using APT representations. We see that the results using standard PPMI ($\alpha = 1$) significantly outperform the result reported in Reddy et al. (2011), which demonstrates the superiority of a typed dependency space over an untyped dependency space. Smoothing the PPMI calculation with a value of $\alpha = 0.75$ generally has a further small positive effect. On average, the results when probabilities are composed and PPMI is calculated as part of the similarity calculation (`CF`) are slightly higher than the results when PPMI weights are composed (`CS`). Regarding different composition operations, $\bigsqcup_{\text{UNI}}$ generally outperforms $\bigsqcup_{\text{INT}}$. In general, the unaligned model outperforms the aligned model. However, a small but statistically significant performance gain is generally made using the hybrid model. Therefore aligned APT composition and unaligned APT composition are predicting different contexts for compound nouns which all contribute to a better estimate of the compositionality of the phrase.

5 Conclusions and further work

We have shown that combining traditional compositional methods with state-of-the-art low-dimensional word representations can improve results over the state-of-the-art. Further improvements can be achieved using an integrated compositional distributional approach based on APT representations. This approach maintains syntactic structure within the contextual features of words which is then central to the compositional process. We argue that some knowledge of syntactic structure is crucial in the fine-grained understanding of language. Since compositionality detection also provides a way of evaluating compositional methods without confounding judgements of phrase similarity with judgements of compositionality, it appears that the APT approach to composition is reasonably promising. Further work is of course needed with other datasets and other

[4] Across all models, optimal values were in the range [0.3,0.5].

types of phrase. For example, it would be interesting to apply these models in German and evaluate their performance on a German noun-noun compound compositionality dataset (Schulte im Walde et al., 2013; Schulte im Walde et al., 2016).

References

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, December.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193, Cambridge, MA, October. Association for Computational Linguistics.

Georgiana Dinu, Nghia The Pham, and Marco Baroni. 2013. General estimation and evaluation of compositional distributional semantic models. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 50–58, Sofia, Bulgaria, August. Association for Computational Linguistics.

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 WAC4 Workshop at LREC*.

Zellig Harris. 1954. Distributional structure. *Word*, 10:146–162.

Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. 2012. An unsupervised ranking model for noun-noun compositionality. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 132–141, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

Douwe Kiela and Stephen Clark. 2013. Detecting compositionality of multi-word expressions using nearest neighbours in vector space models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1427–1432, Seattle, Washington, USA, October. Association for Computational Linguistics.

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

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

Dmitrijs Milajevs, Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, and Matthew Purver. 2014. Evaluating neural word representations in tensor-based compositional settings. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 708–719, Doha, Qatar, October. Association for Computational Linguistics.

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

Joakim Nivre. 2004. Incrementality in deterministic dependency parsing. In *Proceedings of the ACL Workshop on Incremental Parsing*, pages 50–57.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199, June.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, pages 1–15, Mexico City, Mexico.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 977–983, Denver, Colorado, May–June. Association for Computational Linguistics.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Sabine Schulte im Walde, Stefan Muller, and Stephan Roller. 2013. Exploring vector space models to predict the compositionality of German noun-noun compounds. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics (*SEM)*, Atlanta, USA, June.

Sabine Schulte im Walde, Anna Hatty, Stefan Bott, and Nana Khvtisavrishvili. 2016. Ghost-NN: A representative gold standard of German noun-noun compounds. In *Proceedings of the 10th Conference on Language Resources and Evaluation (LREC)*, Portoroz, Slovenia, May.

Julie Weeds, David Weir, and Jeremy Reffin. 2014. Distributional composition using higher-order dependency vectors. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, pages 11–20, Gothenburg, Sweden, April. Association for Computational Linguistics.

David Weir, Julie Weeds, Jeremy Reffin, and Thomas Kober. 2016. Aligning packed dependency trees: a theory of composition for distributional semantics. *Computational Linguistics, special issue on Formal Distributional Semantics*, 42(4):727–761, December.

Majid Yazdani, Meghdad Farahmand, and James Henderson. 2015. Learning semantic composition to detect non-compositionality of multiword expressions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1733–1742, Lisbon, Portugal, September. Association for Computational Linguistics.