Detecting Compositionality of Multi-Word Expressions using Nearest Neighbours in Vector Space Models

Douwe Kiela
University of Cambridge
Computer Laboratory
douwe.kiela@cl.cam.ac.uk

Stephen Clark
University of Cambridge
Computer Laboratory
stephen.clark@cl.cam.ac.uk

Abstract

We present a novel unsupervised approach to detecting the compositionality of multi-word expressions. We compute the compositionality of a phrase through substituting the constituent words with their "neighbours" in a semantic vector space and averaging over the distance between the original phrase and the substituted neighbour phrases. Several methods of obtaining neighbours are presented. The results are compared to existing supervised results and achieve state-of-the-art performance on a verb-object dataset of human compositionality ratings.

1 Introduction

Multi-word expressions (MWEs) are defined as "idiosyncratic interpretations that cross word boundaries" (Sag et al., 2002). They tend to have a standard syntactic structure but are often semantically non-compositional; i.e. their meaning is not fully determined by their syntactic structure and the meanings of their constituents. A classic example is *kick the bucket*, which means *to die* rather than *to hit a bucket with the foot*. These types of expressions account for a large proportion of day-to-day language interactions (Schuler and Joshi, 2011) and present a significant problem for natural language processing systems (Sag et al., 2002).

This paper presents a novel unsupervised approach to detecting the compositionality of MWEs, specifically of verb-noun collocations. The idea is that we can recognize compositional phrases by substituting related words for constituent words in the phrase: if the result of a substitution yields a meaningful phrase, its individual constituents are likely to contribute toward the overall meaning of the phrase. Conversely, if a substitution yields a non-sensical phrase, its constituents are likely to contribute less or not at all to the overall meaning of the phrase. For the phrase *eat her hat*, for example, we might consider the following substituted phrases:

1. *consume* her hat
2. eat her *trousers*

Both phrases are semantically anomalous, implying that *eat hat* is a highly non-compositional verb-noun collocation. Following a similar procedure for *eat apple*, however, would not lead to an anomaly: *consume apple* and *eat pear* are perfectly meaningful, leading us to believe that *eat apple* is compositional.

In the context of distributional models, this idea can be formalised in terms of vector spaces:

> the average distance between a phrase vector and its substituted phrase vectors is related to its compositionality.

Since we are relying on the relative distances of *phrases* in semantic space, we require a method for computing vectors for phrases. We experimented with a number of composition operators from Mitchell and Lapata (2010), in order to compose constituent word vectors into phrase vectors. The relation between phrase vectors and substituted phrase vectors is most pronounced in the case of

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1427–1432,
Seattle, Washington, USA, 18-21 October 2013.

pointwise multiplication, which has the effect of placing semantically anomalous phrases relatively close together in space (since the vectors for the constituent words have little in common), whereas the semantically meaningful phrases are further apart. This implies that compositional phrases are less similar to their neighbours, which is to say that the greater the average distance between a phrase vector and its substituted phrase vectors, the greater its compositionality.

The contribution of this short focused research paper is a novel approach to detecting the compositionality of multi-word expressions that makes full use of the ability of semantic vector space models to calculate distances between words and phrases. Using this unsupervised approach, we achieve state-of-the-art performance in a direct comparison with existing supervised methods.

2 Dataset and Vectors

The verb-noun collocation dataset from Venkatapathy and Joshi (2005), which consists of 765 verb-object pairs with human compositionality ratings, was used for evaluation. Venkatapathy & Joshi used a support vector machine (SVM) to obtain a Spearman ρ_s correlation of 0.448. They employed a variety of features ranging from frequency to LSA-derived similarity measures and used 10% of the dataset as training data with tenfold cross-validation. McCarthy et al. (2007) used the same dataset and expanded on the original approach by adding WordNet and distributional prototypes to the SVM, achieving a ρ_s correlation of 0.454.

The distributional vectors for our experiments were constructed from the ukWaC corpus (Baroni et al., 2009). Vectors were obtained using a standard window method (with a window size of 5) and the 50,000 most frequent context words as features, with stopwords removed. We also experimented with syntax-based co-occurrence features extracted from a dependency-parsed version of ukWaC, but in agreement with results obtained by Schulte im Walde et al. (2013) for predicting compositionality in German, the window-based co-occurrence method produced better results.

We tried several weighting schemes from the literature, such as t-test (Curran, 2004), positive mutual information (Bullinaria and Levy, 2012) and the ratio of the probability of the context word given the target word[1] to the context word's overall probability (Mitchell and Lapata, 2010). We found that a tf-idf variant called LTU yielded the best results, defined as follows (Reed et al., 2006):

$$w_{ij} = \frac{(\log(f_{ij}) + 1.0)\ \log(\frac{N}{n_j})}{0.8 + 0.2 \times \frac{|\textit{context word}|}{|\textit{avg context word}|}}$$

where f_{ij} is the number of times that the target word and context word co-occur in the same window, n_j is the context word frequency, N is the total frequency and $|\textit{context word}|$ is the total number of occurrences of a context word. Distance is calculated using the standard cosine measure:

$$dist(v_1, v_2) = 1 - \frac{v_1 \cdot v_2}{|v_1||v_2|}$$

where v_1 and v_2 are vectors in the semantic vector space model.

3 Finding Neighbours and Computing Compositionality

We experimented with two different ways of obtaining neighbours for the constituent words in a phrase. Since vector space models lend themselves naturally to similarity computations, one way to get neighbours is to take the k-most similar vectors from a similarity matrix. This approach is straightforward, but has some potential drawbacks: it assumes that we have a large number of vectors to select neighbours from, and becomes computationally expensive when the number of neighbours is increased.

An alternative source for obtaining neighbours is the lexical database WordNet (Fellbaum, 1998). We define neighbours as siblings in the hypernym hierarchy, so that the neighbours of a word can be found by taking the hyponyms of its hypernyms. WordNet also allows us to extract only neighbours of the same grammatical type (yielding noun neighbours for nouns and verb neighbours for verbs, for example). Since not every word has the same number of neighbours in WordNet, we use only the first k

[1] We use *target word* to refer to the word for which a vector is being constructed.

neighbours, which means that the neighbours have to be ranked. An obvious ranking method is to use the frequency with which each neighbour co-occurs with the other constituent(s) of the same phrase. For example, for all the WordNet neighbours of *eat* (for all senses of *eat*), we count the co-occurrences with *hat* in a given window size and rank them accordingly. This ranking method also has the desirable side-effect of performing some word sense disambiguation, at least in some cases. For example, the highly ranked neighbours of *apple* for *eat apple* are likely to be items of food, and not (inedible) trees (apple is also a tree in WordNet).

In order to obtain frequency-ranked neighbours, we used the ukWaC corpus with a window size of 5. One reason for having multiple neighbours is that it allows us to correct for word sense disambiguation errors (as mentioned above), since averaging over results for several neighbours reduces the impact of including incorrect senses. For example, the first 20 neighbours of *eat*, ranked by co-occurrence frequency with all the objects of *eat* in the dataset, are:

> *eat use consume drink sample smoke swallow spend break hit save afford burn partake dine breakfast worry damage deplete drug*

One problem with the evaluation dataset is that it does not solely consist of verb-noun pairs: 84 phrases contain pronouns, while there are also several examples containing words that WordNet considers to be adjectives rather than nouns. This problem was mitigated by part-of-speech tagging the dataset. As neighbours for pronouns (which are not included in WordNet), we used the other pronouns present in the dataset. For the remaining words, we included the part-of-speech when looking up the word in WordNet.

3.1 Average distance compositionality score

We considered several different ways of constructing phrasal vectors. We chose not to use the compositional models of Baroni and Zamparelli (2010) and Socher et al. (2011) because we believe that it is important that our methods are completely unsupervised and do not require any initial learning phase. Hence, we experimented with different ways of constructing phrasal vectors according to Mitchell and Lapata (2010) and found that pointwise multiplication $\odot$ worked best in our experiments. Thus, we define the composed vector $\overrightarrow{eat\ hat}$ as:

$$\overrightarrow{eat} \odot \overrightarrow{hat}$$

We can now compute a compositionality score s_c by averaging the distance between the original phrase vector and its substituted neighbour phrase vectors via the following formula:

$$s_c(\overrightarrow{eat\ hat}) = \frac{1}{2k}(\sum_{i=1}^{k} dist(\overrightarrow{eat} \odot \overrightarrow{hat}, \overrightarrow{eat} \odot \overrightarrow{neighbour_i}) + \sum_{j=1}^{k} dist(\overrightarrow{eat} \odot \overrightarrow{hat}, \overrightarrow{neighbour_j} \odot \overrightarrow{hat}))$$

We also experimented with substituting only for the noun or the verb, and in fact found that only taking neighbours for the verb yields better results:

$$s_c(\overrightarrow{eat\ hat}) = \frac{1}{k} \sum_{j=1}^{k} dist(\overrightarrow{eat} \odot \overrightarrow{hat}, \overrightarrow{neighbour_j} \odot \overrightarrow{hat})$$

To illustrate the method, consider the collocations *take breath* and *lend money*. The annotators assigned these phrases a compositionality score of 1 out of 6 and 6 out of 6, respectively, meaning that the former is non-compositional and the latter is compositional. The distances between the first ten verb-substituted phrases and the original phrase, together with the average distance, are shown in Table 1 and Table 2.

Substituting the verb in the non-compositional phrase yields semantically anomalous vectors, which leads to very small changes in the distance between it and the original phrase vector. This is a result of using pointwise multiplication, where overlapping components are stressed: since the vectors for *take* and *breath* have little overlap outside of

Neighbour	Dist
get breath	0.049
find breath	0.051
use breath	0.050
work breath	0.060
hold breath	0.094
run breath	0.079
carry breath	0.076
look breath	0.065
play breath	0.071
buy breath	0.100
AvgDist	0.069

Table 1: Example *take breath*

Neighbour	Dist
pay money	0.446
put money	0.432
bring money	0.405
provide money	0.442
owe money	0.559
sell money	0.404
cost money	0.482
look money	0.425
distribute money	0.544
offer money	0.428
AvgDist	0.457

Table 2: Example *lend money*

the idiomatic sense in *take breath*, its neighbour-substituted phrases also have little overlap, resulting in a smaller change in distance upon substitution. Conversely, substituting the verb in the compositional phrase yields meaningful vectors, putting them in locations in semantic vector space which are sufficiently far apart to distinguish them from the non-compositional cases.

4 Results

Results are given for the two methods of obtaining neighbours: via frequency-ranked WordNet neighbours and via vector space neighbours. The compositionality score was computed by using only the verb, only the noun, or both constituent neighbours in the substituted phrase vectors.

System	ρ_s
Venkatapathy and Joshi (2005)	0.447
McCarthy et al. (2007)	0.454
AvgDist VSM neighbours-both	0.131
AvgDist VSM neighbours-verb	0.420
AvgDist VSM neighbours-noun	0.245
AvgDist WN-ranked neighbours-both	0.165
AvgDist WN-ranked neighbours-verb	**0.461**
AvgDist WN-ranked neighbours-noun	0.169

Table 3: Spearman ρ_s results

The results are compared with the scores reported in Venkatapathy and Joshi (2005) and McCarthy et al. (2007), which were achieved using SVMs with a wide variety of features. Values of $1 \leq k \leq 20$ were tried. If a phrase has fewer than k neighbours because not enough neighbours have been found to co-occur with the other constituent, we use all of them. The results for $k = 20$ are reported here because that gave the best overall score. The dataset has an inter-annotator agreement of Kendall's τ of 0.61 and a Spearman ρ_s of 0.71 and all reported differences in values are highly significant. Table 3 gives the results.

Note that, even though the current approach is unsupervised (in terms of not having access to compositionality ratings during training, although it does rely on WordNet), it outperforms SVMs that require an ensemble of complex feature sets (some of which are also based on WordNet).

It is interesting to observe that the state-of-the-art performance is reached when only using the verb's neighbours to compute substituted phrase vectors. One might initially expect this not to be the case, since e.g. *eat trousers*, where the noun has been substituted, does not make a lot of sense either — which we would expect to be informative for determining compositionality. There are two possible explanations for this, which might be at play simultaneously: since our dataset consists of verb-object pairs, the verb constituent is always the head word of the phrase, and the dataset contains several so-called "light verbs", which have little semantic content of their own. Head words have been found to have a higher impact on compositionality scores for compound nouns: Reddy et al. (2011) weighted

the contribution of individual constituents in such a way that the modifier's contribution is included but is weighted less highly than the head's contribution, which led to an improvement in performance. Our results might be improved by weighting the contribution of constituent words in a similar fashion, and by more closely examining the impact of light verbs for the compositionality of a phrase.

5 Related Work

The past decade has seen extensive work on computational and statistical methods in detecting the compositionality of MWEs (Lin, 1999; Schone and Jurafsky, 2001; Katz and Giesbrecht, 2006; Sporleder and Li, 2009; Biemann and Giesbrecht, 2011). Many of these methods rely on distributional models and vector space models (Schütze, 1993; Turney and Pantel, 2010; Erk, 2012). Work has been done on different types of phrases, including work on particle verbs (McCarthy et al., 2003; Bannard et al., 2003), verb-noun collocations (Venkatapathy and Joshi, 2005; McCarthy et al., 2007), adjective-noun combinations (Vecchi et al., 2011) and noun-noun compounds (Reddy et al., 2011), as well as on languages other than English (Schulte im Walde et al., 2013). Recent developments in distributional compositional models (Widdows, 2008; Mitchell and Lapata, 2010; Baroni and Zamparelli, 2010; Coecke et al., 2010; Socher et al., 2011) have opened up a number of possibilities for constructing vectors for phrases, which have also been applied to compositionality tests (Giesbrecht, 2009; Kochmar and Briscoe, 2013).

This paper takes that work a step further: by constructing phrase vectors and evaluating these vectors on a dataset of human compositionality ratings, we show that existing compositional models allow us to detect compositionality of multi-word expressions in a straightforward and intuitive manner.

6 Conclusion

We have presented a novel unsupervised approach that can be used to detect the compositionality of multi-word expressions. Our results show that the underlying intuition appears to be sound: substituting neighbours may lead to meaningful or meaningless phrases depending on whether or not the phrase is compositional. This can be formalized in vector space models to obtain compositionality scores by computing the average distance to the original phrase's substituted neighbour phrases. In this short focused research paper, we show that, depending on how we obtain neighbours, we are able to achieve a higher performance than that achieved by supervised methods which rely on a complex feature set and support vector machines.

Acknowledgments

This work has been supported by EPSRC grant EP/I037512/1. The authors would like to thank Diana McCarthy for providing the dataset; and Ed Grefenstette, Eva Maria Vecchi, Laura Rimell and Tamara Polajnar and the anonymous reviewers for their helpful comments.

References

Colin Bannard, Timothy Baldwin, and Alex Lascarides. 2003. A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL 2003 Workshop on Multiword expressions: analysis, acquisition and treatment*, MWE 03.

M. Baroni and R. Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1183–1193.

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

Chris Biemann and Eugenie Giesbrecht. 2011. Disco-11: Proceedings of the workshop on distributional semantics and compositionality.

John A. Bullinaria and Joseph P. Levy. 2012. Extracting Semantic Representations from Word Co-occurrence Statistics: Stop-lists, Stemming and SVD. *Behavior Research Methods*, 44:890–907.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. In J. van Bentham, M. Moortgat, and W. Buszkowski, editors, *Linguistic Analysis (Lambek Festschrift)*, volume 36, pages 345–384.

James Curran. 2004. *From Distributional to Semantic Similarity*. Ph.D. thesis, University of Edinburgh.

Katrin Erk. 2012. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Eugenie Giesbrecht. 2009. In search of semantic compositionality in vector spaces. In Sebastian Rudolph, Frithjof Dau, and SergeiO. Kuznetsov, editors, *Conceptual Structures: Leveraging Semantic Technologies*, volume 5662 of *Lecture Notes in Computer Science*, pages 173–184. Springer Berlin Heidelberg.

Graham Katz and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, MWE '06, pages 12–19.

Ekaterina Kochmar and Ted Briscoe. 2013. Capturing Anomalies in the Choice of Content Words in Compositional Distributional Semantic Space. In *Recent Advances in Natural Language Processing*.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 317–324.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment - Volume 18*, MWE '03, pages 73–80.

Diana McCarthy, Sriram Venkatapathy, and Aravind Joshi. 2007. Detecting compositionality of verb-object combinations using selectional preferences. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 369–379.

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

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of The 5th International Joint Conference on Natural Language Processing 2011 (IJCNLP 2011), Thailand*.

J.W. Reed, Y. Jiao, T.E. Potok, B.A. Klump, M.T. Elmore, and A.R. Hurson. 2006. TF-ICF: A new term weighting scheme for clustering dynamic data streams. In *Machine Learning and Applications, 2006. ICMLA '06. 5th International Conference on*, pages 258–263.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword expressions: A Pain in the Neck for NLP. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, pages 1–15.

Patrick Schone and Daniel Jurafsky. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of Empirical Methods in Natural Language Processing*, EMNLP '01.

William Schuler and Aravind K. Joshi. 2011. Tree-rewriting models of multi-word expressions. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, MWE '11, pages 25–30.

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 2nd Joint Conference on Lexical and Computational Semantics*, pages 255–265, Atlanta, GA.

Hinrich Schütze. 1993. Word space. In *Advances in Neural Information Processing Systems 5*, pages 895–902. Morgan Kaufmann.

Richard Socher, Cliff Lin, Andrew Y. Ng, and Christopher D. Manning. 2011. Parsing Natural Scenes and Natural Language with Recursive Neural Networks. In *The 28th International Conference on Machine Learning*, ICML 2011.

Caroline Sporleder and Linlin Li. 2009. 2009. unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, EACL '09.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: vector space models of semantics. *J. Artif. Int. Res.*, 37(1):141–188, January.

Eva Maria Vecchi, Marco Baroni, and Roberto Zamparelli. 2011. (linear) maps of the impossible: Capturing semantic anomalies in distributional space. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pages 1–9, Portland, Oregon, USA, June. Association for Computational Linguistics.

Sriram Venkatapathy and Aravind K. Joshi. 2005. Measuring the relative compositionality of verb-noun (v-n) collocations by integrating features. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 899–906.

Dominic Widdows. 2008. Semantic vector products: Some initial investigations. In *Second AAAI Symposium on Quantum Interaction*, Oxford.