Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Geometry of Compositionality

Hongyu Gong, Suma Bhat, Pramod Viswanath
hgong6@illinois.edu, spbhat2@illinois.edu, pramodv@illinois.edu
Department of Electrical and Computer Engineering
University of Illinois at Urbana Champaign, USA

Abstract

This paper proposes a simple test for compositionality (i.e., literal usage) of a word or phrase in a *context-specific* way. The test is computationally simple, relying on no external resources and only uses a set of trained word vectors. Experiments show that the proposed method is competitive with state of the art and displays high accuracy in context-specific compositionality detection of a variety of natural language phenomena (idiomaticity, sarcasm, metaphor) for different datasets in multiple languages. The key insight is to connect compositionality to a curious geometric property of word embeddings, which is of independent interest.

1 Introduction

Idiomatic expressions and figurative speech are key components of the creative process that embodies natural language. One expression type is multiword expressions (MWEs) – phrases with semantic idiosyncrasies that cross word boundaries (Sag et al. 2002). Examples of MWEs include *by and large*, *spill the beans* and *part of speech*. As such, these phrases are idiomatic, in that their meanings cannot be inferred from the meaning of their component words, and are hence termed *non-compositional* phrases as opposed to being *compositional* phrases.

A particularly intriguing aspect of MWEs is their ability to take on degrees of compositionality depending on the context they are in. For example, consider two contexts in which the phrase *bad egg* occurs.
(1) Ensure that one **bad egg** doesn't spoil good businesses for those that care for their clientele.
(2) I don't know which hen is laying the **bad egg** but when I crack it, it explodes! It is all creamy yellowish with very little odor.
In (1), the phrase has a non-compositional interpretation to mean 'an unpleasant person', whereas in (2), the phrase has the meaning of a noun phrase whose head is egg and modifier is bad. This context-dependent degree of compositionality of an MWE poses significant challenges to natural language processing applications. In machine translation, instead of processing the MWE as a whole, a literal translation of its components could result in a meaningless phrase in the target language, e.g., *chemin de fer* from French to English to be *way of iron* in place of *railway* (Bouamor, Semmar, and Zweigenbaum 2012). In information retrieval, the retrieved document matching a component word is irrelevant given the meaning of the MWE *hot dog*. Hence, identifying the compositionality of MWEs is an important subtask in a computational system.

As another example, consider the word *love* in the following two contexts. In the first: "I **love** going to the dentist. Been waiting for it all week!", the word has a non-literal (hence non-compositional) and *sarcastic* interpretation to actually mean the exact opposite of the literal (compositional) sense, which is to "like". In the second: "I **love** strawberry ice cream; it's simply my favorite", the same word has the compositional meaning. Again, the degree of compositionality is crucially context-dependent.

Yet another example of compositionality involves *metaphors*. Consider the word *angel* in the following two contexts:
(1) The girl is an **angel**; she is helpful to the children.
(2) The **angels** are sure keeping busy, what with all his distractions and mishaps.
In (1) the word has a figurative sense (i.e., non-compositional interpretation) whereas in (2), the word has the compositional meaning of a "divine being". Again, the degree of compositionality is crucially context-dependent.

In this paper our focus is to decide the compositionality of a word or a phrase using its *local linguistic context*. Our approach only relies on the use of word embeddings, which capture the "meaning" of a word using a low-dimensional vector. Our compositionality prediction algorithm brings two key innovations: (1) It leverages the crucial contextual information that dictates the compositionality of a phrase or word; (2) The prediction mechanism is completely independent of external linguistic resources. Both these are significant improvements over recent works with similar goals: compositionality of MWEs (Salehi, Cook, and Baldwin 2015), works on sarcasm (Wallace et al. 2014) and metaphor detection (Tsvetkov et al. 2014) (the latter works rely significantly on external linguistic resources and access to labeled training data).

To the best of our knowledge, this is the first unsupervised study on *context-dependent* phrase compositionality and the first resource-independent study on sarcasm and metaphor

identification. This work is centered around two primary questions:
(1) How can the semantics of a long context be represented by word embeddings?
(2) How can we decide the compositionality of a phrase based on its embeddings and that of its context?

We answer these questions by connecting the notion of compositionality to a geometric property of word embeddings. The key insight is that the context word vectors (suitably compressed) reside roughly in a low dimensional *linear* subspace and compositionality turns out to be related to the projection of the word/phrase embeddings (suitably compresed to a single vector) onto this context subspace.

The key justification for our approach comes from empirical results that outperform state of the art methods on many metrics, while being competitive on the others. We use three standard datasets spanning two MWE construction types (noun compounds and verb particle constructions) in two languages (English and German) in addition to a dataset in Chinese (heretofore unexplored language), and standard datasets for detection of metaphor and sarcasm in addition to a new dataset for sarcasm detection from Twitter. We summarize our contributions below.

Compositional Geometry: We show that a word (or MWE) and its context are geometrically related as jointly lying in a *linear* subspace, when it appears in a compositional sense, but not otherwise.

Compositionality decision: The *only* input to the algorithm is a set of trained word vectors after the preprocessing step of removing function words and the algorithm perform simple PCA (principal component analysis) operations.

Multi-lingual applicability: The algorithm is very general, relies on no external resources and is agnostic to the specifics of one language; we demonstrate strong test results across different languages.

We begin next with a discussion of the geometry of compositionality leading directly to our context-based algorithm for compositionality detection. The test is competitive with or superior to state of the art in a variety of contexts and languages and in various metrics.

2 Compositionality and the Geometry of Word Embeddings

Our goal is to detect the compositionality level of a given occurrence of a *word/phrase* within a *sentence* (the context). Our **main contribution** is the discovery of a geometric property of vector embeddings of context words (excluding the function words) within a sentence: they roughly occupy a low dimensional linear subspace which can be empirically extracted via a standard technique: principal component analysis (PCA) (Shlens 2014).

We stack the d-dimension vectors $v_1, \ldots, v_n$, corresponding to n words in a sentence, to form a $d \times n$ matrix X. PCA finds a $d \times m$ $(m < n)$ matrix X' which maximizes the data variance with reduced dimension. Here X' consists of m new vectors, $v'_1, ..., v'_m$. Now the original data X is represented by fewer vectors of X', where vectors v and v' are d-dimensional (m is chosen such that a large enough fraction – a hyperparameter – of the variance of X is captured in X').

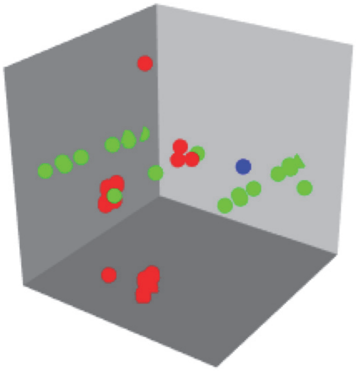


Figure 1: Geometry of phrase and context. The compositional context embeddings of *cutting edge* are denoted by green points, and the non-compositional context embeddings by red points. The embedding of phrase *cutting edge* is denoted by the blue point. Note that the phrase embedding is very close to the space of the compositional context while being farther from the space of its non-compositional context.

When the phrase of interest occurs in a compositional sense, then the phrase's compositional embedding is roughly close to the subspace associated with the context embeddings (extracted using PCA from context words). Intuitively this happens because compositionality is tantamount to individual words themselves being directly related (i.e., occur together often enough) to (a majority of) the context words.

We illustrate this phenomenon via an example found in Table 1. Consider the phrase "cutting edge". When words like *sharp, side* and *tool* appear in the context, "cutting edge" tends to have its compositional meaning. Conversely, when words like *productions, technology* and *competitive* are in the context, "cutting edge" is more likely to be an idiom. We project the embeddings of the phrase and the two contexts to three-dimensions to visualize the geometric relationship, cf. Figure 1.

It is immediate that the phrase embedding occupies the same subspace as the context when it is used in the compositional sense, while it is far from the subspace of the context when used in the non-compositional sense. The precise formulation of the projection operations used in this illustration is discussed next.

2.1 Compositional Semantic Representation

Suppose a sentence t consists of n content words $\{w_1, ..., w_n\}$ with respective vector embeddings $\{v_1, ..., v_n\}$. Two possible representations of the "meaning" of t are the following:

average vector representation: $v_t = v_1 + ... + v_n$, adding all component word vectors together, as in (Mitchell and Lapata 2010) and several works on phrase2vec (Gershman and Tenenbaum 2015) and sentence2vec (Faruqui et al. 2015).

PCA subspace representation: Denote the word vectors by $X = [v_1, \ldots, v_n]$, and the PCA output $X' = [v'_1, ..., v'_m]$, where v'_i are principal components extracted from X using the PCA operation. Now the sentence t is represented by the (span of columns of) matrix X' instead of a single vector as in average vector representation. Choosing to represent the sentence by multiple vectors is a key innovation of this paper and is fairly critical to the empirical results we demonstrate.

Note that the PCA operation returns a $(d \times m)$ matrix X' and thus PCA is used to reduce the "number of word vectors" instead of the embedding dimension. In our experiments, $d = 200$, $n \approx 10 - 20$, and $m \approx 3$. PCA extracts the most important information conveyed in the sentence with only m vectors. Further, we only take the linear span of the m principal directions (column span of X'), i.e., a *subspace* as the representation of sentence t.

Let p be a single word (in the metaphor and sarcasm settings) or a bigram phrase (in the MWE setting) that we would like to test for compositional use. Suppose that p has a single-vector representation v_p, and context embedding is represented by the subspace S_c spanned by the m vectors $(v'_1, \ldots, v'_m)$. Our test involves projecting the phrase embedding v_p on the context subspace S_c. Denote the orthogonal projection vector by v'_p, where v'_p lies in S_c, and $v'_p = \arg\max_{v \in \mathbb{R}^d} \frac{v^T v_p}{\|v\| \cdot \|v_p\|}$.

Compositionality Score is the cosine distance between v_p and v'_p (the inner product between the vectors normalized by their lengths); this measures the degree to which the word/phrase meaning agrees with its context: *the larger the cosine similarity, the more the compositionality*.

Based on the commonly-used distributional hypothesis: word or phrase meaning can be inferred from its context (Rubenstein and Goodenough 1965), we note that the *local* context (neighboring words) is crucial in deciphering the compositional sense of the word or phrase. This is in contrast to prior works that use the global context (the whole document or corpus) for semantic analysis, without accounting for the context-dependence of polysemy (Reddy, McCarthy, and Manandhar 2011).

At times, the word(s) being tested themselves exhibit polysemous behavior (example: *check* in *blank check*) (Mu, Bhat, and Viswanath 2016). In such cases, it makes sense to consider multiple embeddings for different word senses (we use MSSG representations (Neelakantan et al. 2014)): each word has a single global embedding and two sense embeddings. We propose to use global word embeddings to represent the context, and sense embeddings for phrases semantics, allowing for multiple compositionality scores. We then measure the relevance between a phrase and its context by the *maximum* of the different compositionality scores.

Our compositionality detection algorithm uses only two hyperparameters: *variance ratio* (used to decide the amount of variance PCA should capture) and *threshold* (used to test if the compositionality score is above or below this value). Since compositionality testing is essentially a supervised learning task: in order to provide one of two labels, we need to tune these parameters based on a (gold) training set. We see in the experiment sections that these parameters are robustly trained on small training sets and are fairly invariant in their values across different datasets, languages and tasks (variance ratio equal to about 0.6 generally achieves good performance).

3 MWE Compositionality Detection

We evaluate our context-based compositionality detection method empirically by considering 3 specific, but vastly distinct, tasks: a) Predicting the compositionality of phrases that can have either the idiomatic sense or the literal sense depending on the context (the focus of this section), b) Sarcasm detection at the level of a specific word and at the level of a sentence, and c) Detecting whether a given phrase has been used in its metaphoric sense or literal sense. The latter two tasks are the focus of the next two sections. For each of the tasks we use standard datasets used in state-of-the-art studies, as well as those we specifically constructed for the experiments. We include datasets in German and Chinese in addition to those available in English to highlight the multi-lingual and language-agnostic capabilities of our algorithm.

The training corpus of embeddings in English, Chinese and German are obtained from polyglot (Al-Rfou, Perozzi, and Skiena 2013). Two types of word embeddings are used in the experiments: one trained with CBOW of word2vec (Mikolov et al. 2014), and the other using NP-MSSG of MSSG (Neelakantan et al. 2014).

3.1 Experiment I: Phrase Compositionality

In this part, we evaluate the performance of our algorithm in capturing the semantics of the context and predicting the compositionality of phrases, which we cast as a binary classification task – to decide the phrase compositionality in each context. With reference to the examples in Table 1, the task is to predict that the phrase *cutting edge* is used in its compositional sense in the first instance and a non-compositional one in the second. We perform experiments with different word embeddings (CBOW and MSSG), as well as different composition approximations for both the phrase and the context (average and PCA).

Bi-context Dataset: We construct 2 datasets [1] (one for English and the other for Chinese) consisting of a list of polysemous phrases and their respective contexts (compositional and non-compositional).

The English dataset contains 104 polysemous phrases which are obtained from the idiom dictionary (TheFreeDictionary 2016), and the Chinese dataset consists of 64 phrases obtained from (ChineseDictionary 2016). Their respective contexts are extracted from the corpus provided by polyglot or electronic resources (GoogleBooks 2016). Native English and native Chinese speakers annotated the phrase compositionality for each context.

Detection Results: The results of using both average and PCA subspace representations are shown as accuracy values,

[1] available at: https://github.com/HongyuGong/Geometry-of-Compositionality.git

Phrase	Compositional Context	Non-compositional Context
cutting edge	the flat part of a tool or weapon that (usually) has a **cutting edge**. Edge - a sharp side.	while creating successful film and TV productions, a **cutting edge** artworks collection.
ground floor	Bedroom one with en-suite is on the **ground floor** and has a TV. Furnished with king size bed, two bedside chests of drawers with lamps.	Enter a business organization at the lowest level or from the **ground floor** or to be in a project under-. taking from its inception

Table 1: Examples of English phrases, whose compositionality depends on the context.

	English (CBOW)	English (MSSG)	Chinese (CBOW)	Chinese (MSSG)
avg phrase avg context	80.3	82.7	78.1	50
pca phrase avg context	59.1	70.2	50.7	50.7
avg phrase pca context	82.7	84.6	80.5	75
pca phrase pca context	85.6	**86.1**	81.3	**88.3**

Table 2: Accuracy values (%) for Experiment I: Compositionality detection from contexts.

obtained by comparing the predicted labels with the gold labels provided by human annotators, in Table 2.1. The average vector of all the words (shown to be a very robust sentence representation (Ettinger, Elgohary, and Resnik 2016)) serves as our baseline.

We note that having a PCA approximation for both the phrase and the context, and the use of MSSG embedding yields the best accuracy for this task in both the English and the Chinese datasets; this is an instance where the PCA subspace representation is superior to the average representation. We believe that improving beyond the fairly high accuracy rates is likely to require substantially new ideas as compared to those in this paper.

3.2 Experiment II: Lexical Idiomaticity

Unlike compositionality detection in Experiment I, here we detect component-wise idiomaticity of a two-word phrase in this experiment. For example, "spelling" is literal while "bee" is idiomatic in the phrase "spelling bee". Modifying our method slightly, we take the cosine distance between the embedding of the target word (the first or the second word) and its projection to the space of its context as the measurement of lexical idiomaticity. The smaller the cosine distance, the more idiomatic the component word is. Here we use three datasets available from prior studies for the same task – ENC, EVPC and GNC – and compare our results with the state-of-art in idiomaticity detection.

Dataset: The English noun compounds dataset (ENC), has 90 English noun compounds annotated on a continuous $[0, 5]$ scale for the phrase and component-wise compositionality (Reddy, McCarthy, and Manandhar 2011); the English verb particle constructions (EVPC) contains 160 English verb-particle compounds, whose componentwise compositionality are annotated on a binary scale (Bannard 2006). German noun compounds (GNC), which contains 246 German noun compounds annotated on a continuous [1,7] scale for phrase and component compositionality (Schulte im Walde, Müller, and Roller 2013). In this paper, we cast compositionality prediction as a binary classification task. We set the same threshold of 2.5 to ENC as in (Salehi, Cook, and Baldwin 2014a), a threshold of 4 to GNC and use the binary labels of EVPC. The components with score higher than the threshold are regarded as literal, otherwise, they are idiomatic.

Detection Results: Our subspace-based method (SubSpace) uses CBOW and MSSG embeddings, and we use both average and PCA approximations as context embeddings. Their performance is shown in the row of "SubSpace (CBOW)" and "SubSpace (MSSG)" respectively. We have two baseline methods: (1) PMI: pointwise mutual information. PMI $= \log \frac{P(w_1 w_2)}{P(w_1)P(w_2)}$, where $P(\cdot)$ is the probability of the unigram or bigram (Manning and Schütze 1999). PMI statistically evaluates the cohesion between words. Higher PMI indicates the phrase is more likely to be non-compositional. (2) Average sentence embedding method: while we use PCA, several recent works have shown average word vectors to be robust sentence embeddings (Ettinger, Elgohary, and Resnik 2016) and we measure compositionality by the cosine similarity between the target word vector and the sentence vector. The corresponding performance is reported in the rows of "Avg Cxt (CBOW)" and "Avg Cxt (MSSG)". We only report the best performance of each method in Table 3.

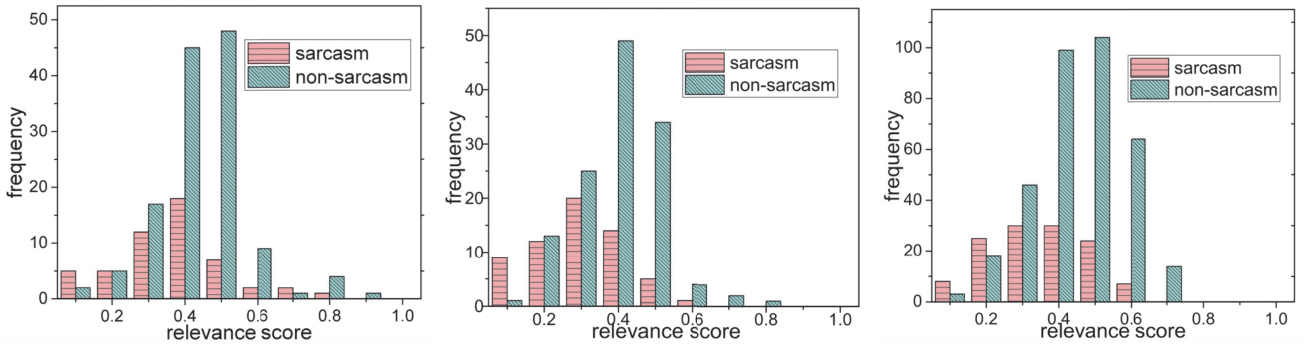
We compare with the state-of-the-art of (Salehi, Cook, and Baldwin 2014a), specifically their methods based on word definitions, synonyms and idiom tags (denoted by ALLDEFS+SYN, ITAG+SYN, ALLDEFS) provided by wikitionary. As we can see from Table 2.1, our method compares favorably to the state-of-art performance while outperforming two baseline methods. The key advantage of our method is our non-reliance on external resources like wikitionary or multilingual translations which are heavily relied upon in the state-of-the-art methods (Salehi, Cook, and Baldwin 2014a; 2014b). Also, unlike the assumption in (Salehi, Cook, and Baldwin 2015), we do not require that the test phrases appear in the embedding training corpus.

4 Sarcasm Detection

Sarcasms, also called irony, are expressions whose actual meaning is quite different - and often opposite to - their literal meaning – and are instances of non-compositional usage (Davidov, Tsur, and Rappoport 2010; Riloff et al. 2013). For

		First Component			Second Component		
Dataset	Method	Precision (%)	Recall (%)	F1 score (%)	Precision (%)	Recall (%)	F1 score (%)
ENC dataset	PMI	50	100	66.7	40.4	100	57.6
	ITAG+SYN	64.5	90.9	75.5	61.8	94.4	**74.7**
	Avg Cxt (MSSG)	68.5	79.5	73.7	61.2	83.3	70.6
	SubSpace (CBOW)	78.4	90.9	**84.2**	67.44	80.6	**73.44**
EVPC dataset	PMI	22.2	68.4	33.5	53.0	80.2	63.8
	ALLDEFS	25.0	97.4	39.8	53.6	97.6	**69.2**
	Avg Cxt (MSSG)	33.8	60.5	43.4	58.0	80.2	67.3
	SubSpace (MSSG)	31.4	86.8	**46.2**	54.4	100	**70.5**
GNC dataset	PMI	44.2	99.0	61.1	26.4	98.4	41.7
	Avg Cxt (CBOW)	45.4	92.6	60.6	29.0	95.4	44.4
	SubSpace (MSSG)	45.5	99.1	**62.4**	30.9	86.2	**45.5**

Table 3: Experiments on ENC, EVPC and GNC Datasets.



(a) word "good"

(b) word "nice"

(c) word "always"

Figure 2: Sarcasm Detection in Tweets

example, the word 'nice' is used in a sarcastic sense in 'It's so nice that a cute video of saving an animal can quickly turn the comments into politcal debates and racist attacks'. The context clues identify sarcasm; in this example, 'nice' is inconsistent with its context words 'debate' and 'attacks'. These ideas are used in prior works to create elaborate features (designed based on a large labeled training set) and build a sarcasm detection system (Ghosh, Guo, and Muresan 2015). Instead, we evaluate our compositionality detection algorithm directly on this task.

Datasets: Tweets are ideal sources of sarcasm datasets. We study words in tweets (a subset of the dataset in (Ghosh, Guo, and Muresan 2015)) that are used both literally and sarcastically (eg: love, like, always) and apply our simple compositionality scoring algorithm. We choose six words "good", "love", "yeah", "glad", "nice" and "always", which have enough occurrences in both literal and sarcastic senses in our downloaded dataset. We apply our relevance scoring algorithm, and count the occurrences in each score bin to see whether our algorithm could distinguish sarcastic usage from literal usage.

The histograms of the compositionality scores for words "good", "nice" and "always" (for sarcastic and literal usages) are plotted in Fig. 2. We can visually see that the two histograms (one for sarcastic usage and the other for literal usage) can be distinguished from each other, for each of these three words. The histogram of sarcastic usage occupies the low-score region peaking in the $[0.3, 0.4)$ bin, while the histogram of literal usage occupies the high-score region with peak in the $[0.4, 0.5)$ bin. This shows that our simple resource-independent compositionality scoring method can distinguish sarcasm and non-sarcasm.

word	'good'	'love'	'yeah'	'nice'	'always'	'glad'
Accuracy	0.744	0.700	0.614	0.763	0.792	0.695
F1 score	0.610	0.64	0.655	0.623	0.605	0.582

Table 4: Twitter Sarcasm Detection

To quantify this extent, we report the accuracy and F1 scores of a simple threshold classifier in each of the six instances in Table 4. We emphasize that this performance is derived for a very small dataset (for each of the words) and is entirely achieved using only a trained set of word vectors – this would be a baseline to build on for the more sophisticated supervised learning systems.

A *quantitative* test is provided via our study on a Reddit irony dataset of 3020 annotated comments (Wallace et al. 2014). An example of an ironic comment is "It's amazing how Democrats view money. It has to come from somewhere you idiots and you signed up to foot the bill. Congratulations." Details of our experiment on the Reddit dataset is available in the full version (Gong, Bhat, and Viswanath

		features	accuracy	f1 score
SVO	state-of-art	279	**0.82**	**0.86**
	SubSpace original sentence	4	0.729	0.744
	SubSpace longer sentence	4	0.809	0.806
AN	state-of-art	360	**0.86**	**0.85**
	SubSpace original sentence	3	0.735	0.744
	SubSpace longer sentence	3	0.80	0.798

Table 5: Metaphor Detection

2016) where our method with its much fewer features gets comparable results to (and in some instances achieve up to a 5% higher F1 score over) the baseline system in (Wallace et al. 2014).

5 Metaphor Detection

Metaphors are usually used to express the abstract sense of a word in noncompositional contexts: in the sentence "Comprehensive solutions marry ideas favored by one party and opposed by the other", the intended meaning of "marry" is "combine", a significant (and figurative) generalization of its literal meaning. As such, metaphors form a key part of noncompositional semantics and are natural targets to study in our generic framework.

Dataset: English datasets consisting of metaphoric and literal uses of two syntactic structures – (subject-verb-object (SVO) and adjective-noun (AN) compounds) – are provided in (Tsvetkov et al. 2014). An example of an SVO metaphor is "Actions talk even louder than phrases", and an example of an AN metaphor is "black humor seems very Irish to me". The SVO dataset contains 111 literal and 111 metaphorical phrases while the AN dataset contains 100 literal and 100 metaphorical phrases.

Algorithm Description: The state-of-the-art work uses training-data-driven feature engineering methods that rely on external resources like WordNet and MRC psycholinguistic database (Tsvetkov et al. 2014). We depart by using scores generated by our compositionality detection algorithm, albeit specific to POS tags (critical for this particular dataset since it is focused on specific syntactic structures), as features for metaphor detection.

For each word in the SVO or AN structure, we obtain a compositionality score with respect to its local context and derive features from these scores: The features we derive for SVO dataset from these scores are: (1) the lowest score in SVO; (2) verb score; (3) ratio between lowest score and highest score; (4) $\min\left(\frac{\text{verb score}}{\text{subj score}}, \frac{\text{subj score}}{\text{verb score}}, \frac{\text{verb score}}{\text{obj score}}, \frac{\text{obj score}}{\text{verb score}}\right)$.

If an SVO phrase is a metaphor, then we expect there will be at least one word which is inconsistent with the context. Thus we include the lowest score as one of the features. Also, the verb score is a feature since verbs are frequently used metaphorically in a phrase. The absolute score is very sensitive to the context, and we also include relative scores to make the features more robust. The relative scores are the ratio between the lowest score and the highest score, and the minimum ratio between verb and subject or object.

The features we get for AN dataset from these scores are: (1) the lowest score in AN; (2) the highest score; (3) ratio between the lowest and the highest score. These features are then fed into a supervised learning system (random forest), analogous to the one in (Tsvetkov et al. 2014) allowing for a fair comparison of the power of the features extracted.

Detection Results: The experiment results on SVO and AN datasets are detailed in Table 5 where the baseline is provided by the results of (Tsvetkov et al. 2014) (which has access to the MRC psycholinguistic database and the supersense corpus). On the full set of original sentences, the performance of our compositionality detection algorithm (with only four features in stark contrast to the more than 100 used in the state of the art) is not too far from the baseline.

Upon a closer look, we find that some of the original sentences are too short, e.g. "The bus eventually arrived". Our context-based method naturally does better with longer sentences and we purified the dataset by replacing sentences whose content words are fewer than 7 with longer sentences extracted from Google Books. We rerun our experiments and the performance on the longer sentences is improved, although it is still a bit below the baseline – again, contrast the very large number of features (extracted using significant external resources) used in the baseline to just 3 or 4 of our approach (extracted in a resource-independent fashion).

6 Related Work

Predicting the compositionality of MWEs is a well recognized challenge with significant recent attention; the proposed approaches can be broadly divided into four categories: using lexical and syntactic properties of MWEs (Pecina and Schlesinger 2006; Cook, Fazly, and Stevenson 2007), distributional hypothesis-based methods (Katz and Giesbrecht 2006), external resource-based methods (Salehi and Cook 2013; Salehi, Cook, and Baldwin 2014b), and word embeddings-based approaches (Salehi, Cook, and Baldwin 2015).

Detection of sarcasm and metaphor is of great importance in language processing applications like text analysis, dialogue systems and sentiment analysis. Recent works have used extensive linguistic resources for sarcasm detection (Ghosh, Guo, and Muresan 2015) and metaphor identification (Tsvetkov et al. 2014). Bringing MWEs, sarcasms and metaphors under a common umbrella of compositionality, followed by a unified framework to study it, is our central contribution. Given that our work is at the intersection of a vast body of topical literature, we have provided a more elaborate discussion of the previous works in this area and comparisons to our work in the full version (Gong, Bhat, and Viswanath 2016).

7 Acknowledgement

This work is supported in part by the IBM C3SR project.

References

Al-Rfou, R.; Perozzi, B.; and Skiena, S. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, 183–192. Sofia, Bulgaria: Association for Computational Linguistics.

Bannard, C. J. 2006. *Acquiring Phrasal Lexicons from Corpora*. Ph.D. Dissertation, University of Edinburgh.

Bouamor, D.; Semmar, N.; and Zweigenbaum, P. 2012. Identifying bilingual multi-word expressions for statistical machine translation. In *LREC*, 674–679.

ChineseDictionary. 2016. Available at: http://www.chinese-dictionary.org. Accessed:2016-05-01.

Cook, P.; Fazly, A.; and Stevenson, S. 2007. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the workshop on a broader perspective on multiword expressions*, 41–48. Association for Computational Linguistics.

Davidov, D.; Tsur, O.; and Rappoport, A. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the fourteenth conference on computational natural language learning*, 107–116. Association for Computational Linguistics.

Ettinger, A.; Elgohary, A.; and Resnik, P. 2016. Probing for semantic evidence of composition by means of simple classification tasks. *the Association for Computational Linguistics* 134.

Faruqui, M.; Dodge, J.; Jauhar, S. K.; Dyer, C.; Hovy, E.; and Smith, N. A. 2015. Retrofitting word vectors to semantic lexicons. Association for Computational Linguistics.

Gershman, S. J., and Tenenbaum, J. B. 2015. Phrase similarity in humans and machines. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*. Citeseer.

Ghosh, D.; Guo, W.; and Muresan, S. 2015. Sarcastic or not: Word embeddings to predict the literal or sarcastic meaning of words. 1003–1012.

Gong, H.; Bhat, S.; and Viswanath, P. 2016. Geometry of compositionality. *arXiv preprint arXiv:1611.09799*.

GoogleBooks. 2016. Available at: https://books.google.com. Accessed: 2016-05-03.

Katz, G., and Giesbrecht, E. 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*, 12–19. Association for Computational Linguistics.

Manning, C., and Schütze, H. 1999. Collocations. *Foundations of statistical natural language processing* 141–77.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2014. word2vec.

Mitchell, J., and Lapata, M. 2010. Composition in distributional models of semantics. *Cognitive science* 34(8):1388–1429.

Mu, J.; Bhat, S.; and Viswanath, P. 2016. Geometry of polysemy. *arXiv preprint arXiv:1610.07569*.

Neelakantan, A.; Shankar, J.; Passos, A.; and McCallum, A. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. *Conference on Empirical Methods in Natural Language Processing*.

Pecina, P., and Schlesinger, P. 2006. Combining association measures for collocation extraction. In *Proceedings of the COLING/ACL on Main conference poster sessions*, 651–658. Association for Computational Linguistics.

Reddy, S.; McCarthy, D.; and Manandhar, S. 2011. An empirical study on compositionality in compound nouns. In *IJCNLP*, 210–218.

Riloff, E.; Qadir, A.; Surve, P.; De Silva, L.; Gilbert, N.; and Huang, R. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *EMNLP*, volume 13, 704–714.

Rubenstein, H., and Goodenough, J. B. 1965. Contextual correlates of synonymy. *Communications of the ACM* 8(10):627–633.

Sag, I. A.; Baldwin, T.; Bond, F.; Copestake, A.; and Flickinger, D. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*. Springer. 1–15.

Salehi, B., and Cook, P. 2013. Predicting the compositionality of multiword expressions using translations in multiple languages. In *Second Joint Conference on Lexical and Computational Semantics*, volume 1, 266–275.

Salehi, B.; Cook, P.; and Baldwin, T. 2014a. Detecting non-compositional mwe components using wiktionary. In *Conference on Empirical Methods in Natural Language Processing*, 1792–1797.

Salehi, B.; Cook, P.; and Baldwin, T. 2014b. Using distributional similarity of multi-way translations to predict multiword expression compositionality. In *EACL*, 472–481.

Salehi, B.; Cook, P.; and Baldwin, T. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *the North American Chapter of the Association for Computational Linguistics*, 977–983.

Schulte im Walde, S.; Müller, S.; and Roller, S. 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*, 255–265.

Shlens, J. 2014. A tutorial on principal component analysis. *CoRR* abs/1404.1100.

TheFreeDictionary. 2016. Available at: http://idioms.thefreedictionary.com. Accessed: 2016-04-20.

Tsvetkov, Y.; Boytsov, L.; Gershman, A.; Nyberg, E.; and Dyer, C. 2014. Metaphor detection with cross-lingual model transfer.

Wallace, B. C.; Do Kook Choe, L. K.; Kertz, L.; and Charniak, E. 2014. Humans require context to infer ironic intent (so computers probably do, too). In *the Association for Computational Linguistics*, 512–516.