An Empirical Study on Compositionality in Compound Nouns

Siva Reddy
University of York, UK
siva@cs.york.ac.uk

Diana McCarthy
Lexical Computing Ltd, UK
diana@dianamccarthy.co.uk

Suresh Manandhar
University of York, UK
suresh@cs.york.ac.uk

Abstract

A multiword is compositional if its meaning can be expressed in terms of the meaning of its constituents. In this paper, we collect and analyse the compositionality judgments for a range of compound nouns using Mechanical Turk. Unlike existing compositionality datasets, our dataset has judgments on the contribution of constituent words as well as judgments for the phrase as a whole. We use this dataset to study the relation between the judgments at constituent level to that for the whole phrase. We then evaluate two different types of distributional models for compositionality detection – constituent based models and composition function based models. Both the models show competitive performance though the composition function based models perform slightly better. In both types, additive models perform better than their multiplicative counterparts.

1 Introduction

Compositionality is a language phenomenon where the meaning of an expression can be expressed in terms of the meaning of its constituents. Multiword expressions (Sag et al., 2002, MWEs) are known to display a continuum of compositionality (McCarthy et al., 2003) where some of them are compositional e.g. "swimming pool", some are non-compositional e.g. "cloud nine", and some in between e.g. "zebra crossing".

The past decade has seen interest in developing computational methods for compositionality in MWEs (Lin, 1999; Schone and Jurafsky, 2001; Baldwin et al., 2003; Bannard et al., 2003; McCarthy et al., 2003; Venkatapathy and Joshi, 2005; Katz and Giesbrecht, 2006; Sporleder and Li, 2009). Recent developments in vector-based semantic composition functions (Mitchell and Lapata, 2008; Widdows, 2008) have also been applied to compositionality detection (Giesbrecht, 2009).

While the existing methods of compositionality detection use constituent word level semantics to compose the semantics of the phrase, the evaluation datasets are not particularly suitable to study the contribution of each constituent word to the semantics of the phrase. Existing datasets (McCarthy et al., 2003; Venkatapathy and Joshi, 2005; Katz and Giesbrecht, 2006; Biemann and Giesbrecht, 2011) only have the compositionality judgment of the whole expression without constituent word level judgment, or they have judgments on the constituents without judgments on the whole (Bannard et al., 2003). Our dataset allows us to examine the relationship between the two rather than assume the nature of it.

In this paper we collect judgments of the contribution of constituent nouns within noun-noun compounds (section 2) alongside judgments of compositionality of the compound. We study the relation between the contribution of the parts with the compositionality of the whole (section 3). We propose various constituent based models (section 4.3) which are intuitive and related to existing models of compositionality detection (section 4.1) and we evaluate these models in comparison to composition function based models. All the models discussed in this paper are built using a distributional word-space model approach (Sahlgren, 2006).

2 Compositionality in Compound Nouns

In this section, we describe the experimental setup for the collecting compositionality judgments of English compound nouns. All the existing datasets focused either on verb-particle, verb-noun or adjective-noun phrases. Instead, we focus on *compound nouns* for which resources are rel-

atively scarce. In this paper, we only deal with compound nouns made up of two words separated by space.

2.1 Annotation setup

In the literature (Nunberg et al., 1994; Baldwin et al., 2003; Fazly et al., 2009), compositionality is discussed in many terms including simple decomposable, semantically analyzable, idiosyncratically decomposable and non-decomposable. For practical NLP purposes, Bannard et al. (2003) adopt a straightforward definition of a compound being compositional if *"the overall semantics of the multi-word expression (here compound) can be composed from the simplex semantics of its parts, as described (explicitly or implicitly) in a finite lexicon"*. We adopt this definition and pose compositionality as a literality issue. *A compound is compositional if its meaning can be understood from the literal (simplex) meaning of its parts*. Similar views of compositionality as literality are found in (Lin, 1999; Katz and Giesbrecht, 2006). In the past there have been arguments in favor/disfavor of compositionality as literality approach (e.g. see (Gibbs, 1989; Titone and Connine, 1999)). The idea of viewing compositionality as literality is also motivated from the shared task organized by Biemann and Giesbrecht (2011). From here on, we use the terms compositionality and literality interchangeably.

We ask humans to score the compositionality of a phrase by asking them *how literal the phrase is*. Since we wish to see in our data the extent that the phrase is compositional, and to what extent that depends on the contribution in meaning of its parts, we also ask them *how literal the use of a component word is within the given phrase*.

For each compound noun, we create three separate tasks – one for each constituent's literality and one for the phrase compositionality. The motivation behind using three separate tasks is to make the scoring mechanism for each task independent of the other tasks. This enables us to study the actual relation between the constituents and the compound scores without any bias to any particular annotator's way of arriving at the scores of the compound w.r.t. the constituents.

There are many factors to consider in eliciting compositionality judgments, such as ambiguity of the expression and individual variation of annotator in background knowledge. To control for this, we ask subjects if they can interpret the meaning of a compound noun from *only* the meaning of the component nouns where we also provide contextual information. All the possible definitions of a compound noun are chosen from WordNet (Fellbaum, 1998), Wiktionary or defined by ourselves if some of the definitions are absent. Five examples of each compound noun are randomly chosen from the ukWaC (Ferraresi et al., 2008) corpus and the same set of examples are displayed to all the annotators. The annotators select the definition of the compound noun which occurs most frequently in the examples and then score the compound for literality based on the most frequent definition.

We have two reasons for making the annotators read the examples, choose the most frequent definition and base literality judgments on the most frequent definition. The first reason is to provide a context to the decisions and reduce the impact of ambiguity. The second is that distributional models are greatly influenced by frequency and since we aim to work with distributional models for compositionality detection we base our findings on the most frequent sense of the compound noun. In this work we consider the compositionality of the noun-noun compound type without token based disambiguation which we leave for future work.

2.2 Compound noun dataset

We could not find any compound noun datasets publicly available which are marked for compositionality judgments. Korkontzelos and Manandhar (2009) prepared a related dataset for compound nouns but compositionality scores were absent and their set contains only 38 compounds. There are datasets for verb-particle (McCarthy et al., 2003), verb-noun judgments (Biemann and Giesbrecht, 2011; Venkatapathy and Joshi, 2005) and adjective-noun (Biemann and Giesbrecht, 2011). Not only are these not the focus of our work, but also we wanted datasets with each constituent word's literality score. Bannard et al. (2003) obtained judgments on whether a verb-particle construction implies the verb or the particle or both. The judgments were binary and not on a scale and there was no judgment of compositionality of the whole construction. Ours is the first attempt to provide a dataset which have both scalar compositionality judgments of the phrase as well as the literality score for each component word.

We aimed for a dataset which would include compound nouns where: 1) both the component words are used literally, 2) the first word is used literally but not the second, 3) the second word is used literally but not the first and 4) both the words are used non-literally. Such a dataset would provide stronger evidence to study the relation between the constituents of the compound noun and its compositionality behaviour.

We used the following heuristics based on WordNet to classify compound nouns into 4 above classes.

1. Each of the component word exists either in the hypernymy hierarchy of the compound noun or in the definition(s) of the compound noun. e.g. *swimming pool* because *swimming* exists in the WordNet definition of *swimming pool* and *pool* exists in the hypernymy hierarchy of *swimming pool*
2. Only the first word exists either in the hypernymy hierarchy or in the definition(s) of the compound and not the second word. e.g. *night owl*
3. Only the second word exists either in the hypernymy hierarchy or in the definition(s) of the compound and not the first word. e.g. *zebra crossing*
4. Neither of the words exist either in hypernymy hierarchy or in the definition(s) of the compound noun. e.g. *smoking gun*

The intuition behind the heuristics is that if a component word is used literally in a compound, it would probably be used in the definition of the compound or may appear in the synset hierarchy of the compound. We changed the constraints, for example decreasing/increasing the depth of the hypernymy hierarchy, and for each class we randomly picked 30 potential candidates by rough manual verification. There were fewer instances in the classes 2 and 4. In order to populate these classes, we selected additional compound nouns from Wiktionary by manually inspecting if they can fall in either class.

These heuristics were only used for obtaining our sample, they were *not* used for categorizing the compound nouns in our study. The compound nouns in all these temporary classes are merged and 90 compound words are selected which have at least 50 instances in the ukWaC corpus. These 90 compound words are chosen for the dataset.

2.3 Annotators

Snow et al. (2008) used Amazon mechanical turk (AMT) for annotating language processing tasks. They found that although an individual turker (annotator) performance was lower compared to an expert, as the number of turkers increases, the quality of the annotated data surpassed expert level quality. We used 30 turkers for annotating each single task and then retained the judgments with sufficient consensus as described in section 2.4.

For each compound noun, 3 types of tasks are created as described above: a judgment on how literal the phrase is and a judgment on how literal each noun is within the compound. For 90 compound nouns, 270 independent tasks are therefore created. Each of these tasks is assigned to 30 annotators. A task is assigned randomly to an annotator by AMT so each annotator may work on only some of the tasks for a given compound.

2.4 Quality of the annotations

Recent studies[1] shows that AMT data is prone to spammers and outliers. We dealt with them in three ways. **a).** We designed a qualification test[2] which provides an annotator with basic training about literality, and they can participate in the annotation task only if they pass the test. **b).** Once all the annotations (90 phrases * 3 tasks/phrase * 30 annotations/task = 8100 annotations) are completed, we calculated the average Spearman correlation score (ρ) of every annotator by correlating their annotation values with every other annotator and taking the average. We discarded the work of annotators whose ρ is negative and accepted all the work of annotators whose ρ is greater than 0.6. **c).** For the other annotators, we accepted their annotation for a task only if their annotation judgment is within the range of ± 1.5 from the task's mean. Table 1 displays AMT statistics. Overall, each annotator on average worked on 53 tasks randomly selected from the set of 270 tasks. This lowers the chance of bias in the data because of any particular annotator.

Spearman correlation scores ρ provide an estimate of annotator agreement. To know the difficulty level of the three types of tasks described in section 2, ρ for each task type is also displayed in

[1] A study on AMT spammers `http://bit.ly/e1IPil`

[2] The qualification test details are provided with the dataset. Please refer to footnote 3.

No. of turkers participated	260	
No. of them qualified	151	
Turkers with $\rho <= 0$	21	
Turkers with $\rho >= 0.6$	81	
No. of annotations rejected	383	
Avg. submit time (sec) per task	30.4	
	highest ρ	avg. ρ
ρ for phrase compositionality	0.741	0.522
ρ for first word's literality	0.758	0.570
ρ for second word's literality	0.812	0.616
ρ for over all three task types	0.788	0.589

Table 1: Amazon Mechanical Turk statistics

Function f	ρ	$\mathbf{R^2}$
ADD	0.966	0.937
MULT	0.965	0.904
COMB	0.971	0.955
WORD1	0.767	0.609
WORD2	0.720	0.508

Table 3: Correlations between functions and phrase compositionality scores

table 1. It is evident that annotators agree more at word level than phrase level annotations.

For each compound, we also studied the distribution of scores around the mean by observing the standard deviation σ. All the compound nouns along with their mean and standard deviations are shown in table 2.

Ideally, if all the annotators agree on a judgment for a given compound or a component word, the deviation should be low. Among the 90 compounds, 15 of them are found to have a deviation $> \pm 1.5$. We used this threshold to signify annotator disagreement. The reason for disagreement could be due to the ambiguity of the compound e.g. *silver screen, brass ring* or due to the subjective differences of opinion between the annotators.

Overall, the inter annotator agreement (ρ) is high and the standard deviation of most tasks is low (except for a few exceptions). So we are confident that the dataset can be used as a reliable gold-standard with which we conduct experiments. The dataset is publicly available for download[3].

[3]Annotation guidelines, Mechanical Turk hits, qualification test, annotators demographic and educational background, and final annotations are downloadable from `http://sivareddy.in/downloads` or `http://www.dianamccarthy.co.uk/downloads.html`

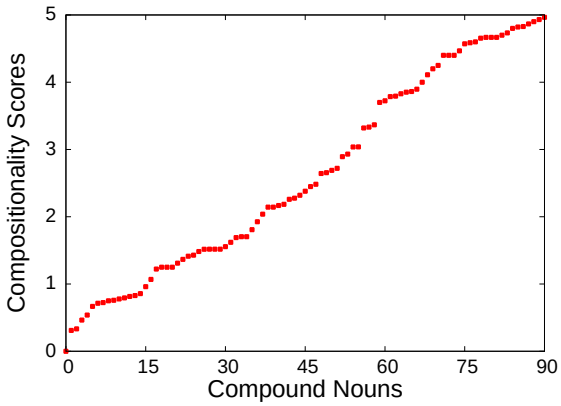

Figure 1: Mean values of phrase-level compositionality scores

3 Analyzing the Human Judgments

By analyzing the mean values of the phrase level annotations, we found that compounds displayed a varied level of compositionality. For some compounds annotators confirm that they can interpret the meaning of a compound from its component words and for some they do not. For others they grade in-between. Figure 1 displays the mean values of compositionality scores of all compounds. Compounds are arranged along the X-axis in increasing order of their score. The graph displays a *continuum of compositionality* (McCarthy et al., 2003). We note that our sample of compounds was selected to exhibit a range of compositionality.

3.1 Relation between the constituents and the phrase compositionality judgments

The dataset allows us to study the relation between constituent word level contributions to the phrase level compositionality scores.

Let $w1$ and $w2$ be the constituent words of the compound $w3$. Let $s1$, $s2$ and $s3$ be the mean literality scores of $w1$, $w2$ and $w3$ respectively. Using a 3-fold cross validation on the annotated data, we tried various function fittings f over the judgments $s1$, $s2$ and $s3$.

- ADD: $a.s1 + b.s2 = s3$
- MULT: $a.s1.s2 = s3$
- COMB: $a.s1 + b.s2 + c.s1.s2 = s3$
- WORD1: $a.s1 = s3$
- WORD2: $a.s2 = s3$

where a, b and c are coefficients.

We performed 3-fold cross validation to evaluate the above functions (two training samples and

Compound	Word1	Word2	Phrase	Compound	Word1	Word2	Phrase
climate change	4.90±0.30	4.83±0.38	4.97±0.18	engine room	4.86±0.34	5.00±0.00	4.93±0.25
graduate student	4.70±0.46	5.00±0.00	4.90±0.30	swimming pool	4.80±0.40	4.90±0.30	4.87±0.34
speed limit	4.93±0.25	4.83±0.38	4.83±0.46	research project	4.90±0.30	4.53±0.96	4.82±0.38
application form	4.77±0.42	4.86±0.34	4.80±0.48	bank account	4.87±0.34	4.83±0.46	4.73±0.44
parking lot	4.83±0.37	4.77±0.50	4.70±0.64	credit card	4.67±0.54	4.90±0.30	4.67±0.70
ground floor	4.66±0.66	4.70±0.78	4.67±0.60	mailing list	4.67±0.54	4.93±0.25	4.67±0.47
call centre	4.73±0.44	4.41±0.72	4.66±0.66	video game	4.50±0.72	5.00±0.00	4.60±0.61
human being	4.86±0.34	4.33±1.14	4.59±0.72	interest rate	4.34±0.99	4.69±0.53	4.57±0.90
radio station	4.66±0.96	4.34±0.80	4.47±0.72	health insurance	4.53±0.88	4.83±0.58	4.40±1.17
law firm	4.72±0.52	3.89±1.50	4.40±0.76	public service	4.67±0.65	4.77±0.62	4.40±0.76
end user	3.87±1.12	4.87±0.34	4.25±0.87	car park	4.90±0.40	4.00±1.10	4.20±1.05
role model	3.55±1.22	4.00±1.03	4.11±1.07	head teacher	2.93±1.51	4.52±1.07	4.00±1.16
fashion plate	4.41±1.07	3.31±2.07	3.90±1.42	balance sheet	3.82±0.89	3.90±0.96	3.86±1.01
china clay	2.00±1.84	4.62±1.00	3.85±1.27	game plan	2.82±1.96	4.86±0.34	3.83±1.23
brick wall	3.16±2.20	3.53±1.86	3.79±1.75	web site	2.68±1.69	3.93±1.18	3.79±1.21
brass ring	3.73±1.95	3.87±1.98	3.72±1.84	case study	3.66±1.12	4.67±0.47	3.70±0.97
polo shirt	1.73±1.41	5.00±0.00	3.37±1.38	rush hour	3.11±1.37	2.86±1.36	3.33±1.27
search engine	4.62±0.96	2.25±1.70	3.32±1.16	cocktail dress	1.40±1.08	5.00±0.00	3.04±1.22
face value	1.39±1.11	4.64±0.81	3.04±0.88	chain reaction	2.41±1.16	4.52±0.72	2.93±1.14
cheat sheet	2.30±1.59	4.00±0.83	2.89±1.11	blame game	4.61±0.67	2.00±1.28	2.72±0.92
fine line	3.17±1.34	2.03±1.52	2.69±1.21	front runner	3.97±0.96	1.29±1.10	2.66±1.32
grandfather clock	0.43±0.78	5.00±0.00	2.64±1.32	lotus position	1.11±1.17	4.78±0.42	2.48±1.22
spelling bee	4.81±0.77	0.52±1.04	2.45±1.25	silver screen	1.41±1.57	3.23±1.45	2.38±1.63
smoking jacket	1.04±0.82	4.90±0.30	2.32±1.29	spinning jenny	4.67±0.54	0.41±0.77	2.28±1.08
number crunching	4.48±0.77	0.97±1.13	2.26±1.00	guilt trip	4.71±0.59	0.86±0.94	2.19±1.16
memory lane	4.75±0.51	0.71±0.80	2.17±1.04	crash course	0.96±0.94	4.23±0.92	2.14±1.27
rock bottom	0.74±0.89	3.80±1.08	2.14±1.19	think tank	3.96±1.06	0.47±0.62	2.04±1.13
night owl	4.47±0.88	0.50±0.82	1.93±1.27	panda car	0.50±0.56	4.66±1.15	1.81±1.07
diamond wedding	1.07±1.29	3.41±1.34	1.70±1.05	firing line	1.61±1.65	1.89±1.50	1.70±1.72
pecking order	0.78±0.92	3.89±1.40	1.69±0.88	lip service	2.03±1.25	1.75±1.40	1.62±1.06
cash cow	4.22±1.07	0.37±0.73	1.56±1.10	graveyard shift	0.38±0.61	4.50±0.72	1.52±1.17
sacred cow	1.93±1.65	0.96±1.72	1.52±1.52	silver spoon	1.59±1.47	1.44±1.77	1.52±1.45
flea market	0.38±0.81	4.71±0.84	1.52±1.13	eye candy	3.83±1.05	0.71±0.75	1.48±1.10
rocket science	0.64±0.97	1.55±1.40	1.43±1.35	couch potato	3.27±1.48	0.34±0.66	1.41±1.03
kangaroo court	0.17±0.37	4.43±1.02	1.37±1.05	snail mail	0.60±0.80	4.59±1.10	1.31±1.02
crocodile tears	0.19±0.47	3.79±1.05	1.25±1.09	cutting edge	0.88±1.19	1.73±1.63	1.25±1.18
zebra crossing	0.76±0.62	4.61±0.86	1.25±1.02	acid test	0.71±1.10	3.90±1.24	1.22±1.26
shrinking violet	2.28±1.44	0.23±0.56	1.07±1.01	sitting duck	1.48±1.48	0.41±0.67	0.96±1.04
rat race	0.25±0.51	2.04±1.32	0.86±0.99	swan song	0.38±0.61	1.11±1.14	0.83±0.91
gold mine	1.38±1.42	0.70±0.81	0.81±0.82	rat run	0.41±0.62	2.33±1.40	0.79±0.66
nest egg	0.79±0.98	0.50±0.87	0.78±0.87	agony aunt	1.86±1.22	0.43±0.56	0.76±0.86
snake oil	0.37±0.55	0.81±1.25	0.75±1.12	monkey business	0.67±1.01	1.85±1.30	0.72±0.69
smoking gun	0.71±0.75	1.00±0.94	0.71±0.84	silver bullet	0.52±1.00	0.55±1.10	0.67±1.15
melting pot	1.00±1.15	0.48±0.63	0.54±0.63	ivory tower	0.38±1.03	0.54±0.68	0.46±0.68
cloud nine	0.47±0.62	0.23±0.42	0.33±0.54	gravy train	0.30±0.46	0.45±0.77	0.31±0.59

Table 2: Compounds with their constituent and phrase level mean±deviation scores

one testing sample at each iteration). The coefficients of the functions are estimated using least-square linear regression technique over the training samples. The average Spearman correlation scores (ρ) over testing samples are displayed in table 3. The goodness of fit R^2 values when trained over the whole data are also displayed in table 3.

Results (both ρ and R^2) clearly show that a relation exists between the constituent literality scores and the phrase compositionality. Existing compositionality approaches on noun-noun compounds such as (Baldwin et al., 2003; Korkontzelos and Manandhar, 2009) use the semantics of only *one* of the constituent words (generally the head word) to determine the compositionality of the phrase. But the goodness of fit R^2 values show that the functions ADD, COMB and MULT which intuitively make use of *both* the constituent scores fit the data better than functions using only one of the constituents. Furthermore, COMB and ADD suggest that additive models are preferable to multiplicative. In this data, the first constituent word plays a slightly more important role than the second in determining compositionality.

Overall, this study suggests that *it is possible to estimate the phrase level compositionality scores given the constituent word level literality scores*. This motivates us to present constituent

based models (section 4.3) for compositionality score estimation of a compound. We begin the next section on computational models with a discussion of related work.

4 Computational Models

4.1 Related work

Most methods in compositionality detection can be classified into two types - those which make use of lexical fixedness and syntactic properties of the MWEs, and those which make use of the semantic similarities between the constituents and the MWE.

Non compositional MWEs are known to have lexical fixedness in which the component words have high statistical association. Some of the methods which exploit this feature are (Lin, 1999; Pedersen, 2011). This property does not hold always because institutionalized MWEs (Sag et al., 2002) are known to have high association even though they are compositional, especially in the case of compound nouns. Another property of non-compositional MWEs is that they show syntactic rigidness which do not allow internal modifiers or morphological variations of the components, or variations that break typical selectional preferences. Methods like (Cook et al., 2007; McCarthy et al., 2007; Fazly et al., 2009) exploit this property. This holds mostly for verbal idioms but not for compound nouns since the variations of any compound noun are highly limited.

Other methods like (Baldwin et al., 2003; Sporleder and Li, 2009) are based on semantic similarities between the constituents and the MWE. Baldwin et al. (2003) use only the information of the semantic similarity between one of the constituents and the compound to determine the compositionality. Sporleder and Li (2009) determine the compositionality of verbal phrases in a given context (token-based disambiguation) based on the lexical chain similarities of the constituents and the context of the MWE. Bannard et al. (2003) and McCarthy et al. (2003) study the compositionality in verb particles and they found that methods based on the similarity between simplex parts (constituents) and the phrases are useful to study semantics of the phrases. These findings motivated our constituent based models along with the findings in section 3.1.

In addition to the constituent based models (section 4.3), there are composition function based vector models (Mitchell and Lapata, 2008; Widdows, 2008) which make use of the semantics of the constituents in a different manner. These models are described in section 4.4 and are evaluated in comparison with the constituent-based models.

The vector space model used in all our experiments is described as follows.

4.2 Vector space model of meaning

Our vector space model is also called a word space model (Sahlgren, 2006, WSM) since we represent a word's meaning in a dimensional space. In the WSM, a word meaning is represented in terms of its Co-occurrences observed in a large corpora where the co-occurrences are stored in a vector format. The lemmatised context words around the target word in a window of size 100 are treated as the co-occurrences. The top 10000 frequent content words in the ukWaC (along with their part-of-speech category) are used for the feature co-occurrences i.e. the dimensionality of the WSM. To measure similarity between two vectors, cosine similarity (sim) is used. Following Mitchell and Lapata (2008), the context words in the vector are set to the ratio of probability of the context word given the target word to the overall probability of the context word[4].

4.3 Constituent based models

Given a compound word $w3$ with the constituents $w1$ and $w2$, constituent based models determine the compositionality score $s3$ of the compound by first determining the literality scores $s1$ and $s2$ of $w1$ and $w2$ respectively (section 4.3.1) and then using one of the functions f (described in section 3.1), the compositionality score $s3$ is estimated using $s3 = f(s1, s2)$ (section 4.3.2).

4.3.1 Literality scores of the constituents

If a constituent word is used literally in a given compound it is highly likely that the compound and the constituent share common co-occurrences. For example, the compound *swimming pool* has the co-occurrences *water*, *fun* and *indoor* which are also commonly found with the constituents *swimming* and *pool*.

We define the literality of a word in a given compound as the similarity between the compound and the constituent co-occurrence vectors i.e. if the number of common co-occurrences are

[4]This is similar to pointwise mutual information without logarithm

numerous then the constituent is more likely to be meant literally in the compound.

Let $v1$, $v2$ and $v3$ be the co-occurrence vectors of $w1$, $w2$ and $w3$. The literality scores $s1$ and $s2$ of $w1$ and $w2$ in the compound $w3$ are defined as

$$s1 = sim(v1, v3)$$
$$s2 = sim(v2, v3)$$

where sim is the cosine similarity between the vectors.

4.3.2 Compositionality of the compound

Given the literality scores $s1$ and $s2$ of the constituents, we can now compute the compositionality score $s3$ of the compound $w3$ using any of the functions f defined in section 3.1.

$$s3 = f(s1, s2)$$

4.4 Composition function based models

In these models (Schone and Jurafsky, 2001; Katz and Giesbrecht, 2006; Giesbrecht, 2009) of compositionality detection, firstly a vector for the compound is composed from its constituents using a compositionality function $\oplus$. Then the similarity between the composed vector and true co-occurrence vector of the compound is measured to determine the compositionality: the higher the similarity, the higher the compositionality of the compound. Guevara (2011) observed that additive models performed well for building composition vectors of phrases from their parts whereas Mitchell and Lapata (2008) found in favor of multiplicative models. We experiment using both the compositionality functions simple addition[5] and simple multiplication, which are the most widely used composition functions, known for their simplicity and good performance.

Vector $\mathbf{v1} \oplus \mathbf{v2}$ for a compound w_3 is composed from its constituent word vectors **v1** and **v2** using the vector addition $a\mathbf{v1} + b\mathbf{v2}$ and simple multiplication **v1v2** where the i^{th} element of $\mathbf{v1} \oplus \mathbf{v2}$ is defined as

$$(a\mathbf{v1} + b\mathbf{v2})_i = a.\mathrm{v1}_i + b.\mathrm{v2}_i$$
$$(\mathbf{v1v2})_i = \mathrm{v1}_i.\mathrm{v2}_i$$

[5]Please note that simple additive model (Mitchell and Lapata, 2008) is different from the additive model described in (Guevara, 2011). In (Mitchell and Lapata, 2008) the coefficients are real numbers whereas in (Guevara, 2011) they are matrices.

	first constituent	second constituent
s1	0.616	–
s2	–	0.707

Table 4: Constituent level correlations

The compositionality score of the compound is then measured using $s3 = sim(\mathbf{v1} \oplus \mathbf{v2}, \mathbf{v3})$ where **v3** is the co-occurrence vector of the compound built from the corpus. For more details of these models please refer to (Mitchell and Lapata, 2008; Giesbrecht, 2009).

4.5 Evaluation

We evaluated all the models on the dataset developed in section 2. Since our dataset has constituent level contributions along with phrase compositionality judgments, we evaluated the constituent based models against both the literality scores of the constituents (section 4.3.1) and the phrase level judgments (section 4.3.2). The composition function models are evaluated only on phrase level scores following (McCarthy et al., 2003; Venkatapathy and Joshi, 2005; Biemann and Giesbrecht, 2011): higher correlation scores indicate better compositionality predictions.

Constituent based models evaluation

Spearman's ρ correlations of s1 and s2 with the human constituent level judgments are shown in table 4. We observed that the predictions for the second constituent are more accurate than those for the first constituent. Perhaps these constitute an easier set of nouns for modelling but we need to investigate this further.

For the phrase compositionality evaluation we did a 3-fold cross validation. The parameters of the functions f (section 4.3.2) are predicted by least square linear regression over the training samples and optimum values are selected. The average Spearman correlation scores of phrase compositionality scores with human judgements on the testing samples are displayed in table 5. The goodness of fit R^2 values when trained over the whole dataset are also displayed.

It is clear that models ADD and COMB which use both the constituents are better predictors of phrase compositionality compared to the single word based predictors WORD1 and WORD2. Both ADD and COMB are competitive in terms of both the correlations (accuracy) and goodness of

Model	ρ	R^2
Constituent Based Models		
ADD	0.686	0.613
MULT	0.670	0.428
COMB	0.682	0.615
WORD1	0.669	0.548
WORD2	0.515	0.410
Compositionality Function Based Models		
$a\mathbf{v1} + b\mathbf{v2}$	0.714	0.620
v1v2	0.650	0.501
RAND	0.002	0.000

Table 5: Phrase level correlations of compositionality scores

fit values. The model MULT shows good correlation but the goodness of fit is lower. First constituent (model WORD1 i.e. $sim(\mathbf{v1}, \mathbf{v3})$) was found to be a better predictor of phrase compositionality than the second (WORD2) following the behaviour of the mechanical turkers as in table 3.

Composition function based models evaluation

These models are evaluated for phrase compositionality scores. As with the constituent based models, for estimating the model parameters a and b of the composition function based models, we did a 3-fold cross validation. The best results of additive model on the training samples are found at a=0.60 and b=0.40. Average Spearman correlation scores of both addition and multiplication models over the testing samples are displayed in table 5. The goodness of fit R^2 values when trained over the whole dataset are also displayed.

Vector addition has a clear upper hand over multiplication in terms of both accuracy and goodness of fit for phrase compositionality prediction.

Winner

For phrase compositionality prediction (table 5), both constituent based and compositionality function based models are found to be competitive, though compositionality function based models perform slightly better. The reason could be because while constituent based models use contextual information of each constituent *independently*, composition function models make use of collective evidence from the contexts of both the constituents *simultaneously*. In the public evaluations of compositionality detection (Biemann and Giesbrecht, 2011), our system (Reddy et al., 2011) which uses the notion of contexts salient to both the constituents achieved better performance than the system which uses only one of the constituent's contexts.

All the results when compared with random baseline (RAND in table 5), which assigns a random compositionality score to a compound, are highly significant.

5 Conclusions

In this paper we examined the compositionality judgments of noun compounds and also the literality judgments of their constituent words. Our study reveals that both the constituent words play a major role in deciding the compositionality of the phrase. We showed that the functions which predict the compositionality using both the constituent literality scores have high correlations with compositionality judgments. Based on this evidence we proposed constituent based models for compositionality detection. We compared constituent based models with compositionality function based models. The additive compositionality functions were slightly superior to the best performing constituent models (again additive) but performance is comparable and we plan to examine more sophisticated constituent models in the future.

All the 8100 annotations collected in this work are released publicly. We hope the dataset can reveal more insights into the compositionality in terms of the contribution from the constituents. Future directions of this work include token based disambiguation of phrases and designing more sophisticated constituent based models. Extending this study on other kinds of phrases such as adjective-noun, verb particle, verb-noun phrases may throw more light into our understanding of compositionality.

Acknowledgements

This work is supported by the European Commission via the EU FP7 INDECT project, Grant No.218086, Research area: SEC-2007-1.2-01 Intelligent Urban Environment Observation System.

References

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*, MWE '03.

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.

Chris Biemann and Eugenie Giesbrecht. 2011. Distributional Semantics and Compositionality 2011: Shared Task Description and Results. In *Proceedings of DISCo-2011 in conjunction with ACL 2011*.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 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*, MWE '07.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computional. Linguististics.*, 35:61–103, March.

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

A. Ferraresi, E. Zanchetta, M. Baroni, and S. Bernardini. 2008. Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the WAC4 Workshop at LREC 2008*.

Raymond W. Gibbs. 1989. Understanding and Literal Meaning. *Cognitive Science*, 13(2):243–251.

Eugenie Giesbrecht. 2009. In Search of Semantic Compositionality in Vector Spaces. In *Proceedings of the 17th International Conference on Conceptual Structures: Conceptual Structures: Leveraging Semantic Technologies*, ICCS '09.

Emiliano Raul Guevara. 2011. Computing Semantic Compositionality in Distributional Semantics. In *Proceedings of the Ninth International Conference on Computational Semantics*, IWCS '2011.

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

Ioannis Korkontzelos and Suresh Manandhar. 2009. Detecting compositionality in multi-word expressions. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the ACL 1999*.

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*, MWE '03.

Diana McCarthy, Sriram Venkatapathy, and Aravind K. Joshi. 2007. Detecting Compositionality of Verb-Object Combinations using Selectional Preferences. In *Proceedings of EMNLP-CoNLL 2007*.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based Models of Semantic Composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio.

Geoffrey Nunberg, Thomas Wasow, and Ivan A. Sag. 1994. Idioms. *Language*, 70(3):491–539.

Ted Pedersen. 2011. Identifying Collocations to Measure Compositionality : Shared Task System Description . In *Proceedings of DISCo-2011 in conjunction with ACL 2011*.

Siva Reddy, Diana McCarthy, Suresh Manandhar, and Spandana Gella. 2011. Exemplar-Based Word-Space Model for Compositionality Detection: Shared Task System Description. In *Proceedings of the ACL Workshop on Distributional Semantics and Compositionality*.

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 CICLing 2002*.

Magnus Sahlgren. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Stockholm University.

Patrick Schone and Daniel Jurafsky. 2001. Is Knowledge-Free Induction of Multiword Unit Dictionary Headwords a Solved Problem? In *Proceedings of EMNLP 2001*.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP 2008*.

Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the EACL 2009*.

Debra A. Titone and Cynthia M. Connine. 1999. On the compositional and noncompositional nature of idiomatic expressions. *Journal of Pragmatics*, 31(12):1655 – 1674. Literal and Figurative Language.

Sriram Venkatapathy and Aravind K. Joshi. 2005. Measuring the relative compositionality of verb-noun (V-N) collocations by integrating features. In *Proceedings of the HLT-EMNLP 2005*.

Dominic Widdows. 2008. Semantic Vector Products: Some Initial Investigations. In *Second AAAI Symposium on Quantum Interaction, Oxford*, March.