An Unsupervised Ranking Model for Noun-Noun Compositionality

Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman
Department of Computer Science
University of Oxford
Wolfson Building, Parks Road
Oxford OX1 3QD, UK
{karl.moritz.hermann,phil.blunsom,stephen.pulman}@cs.ox.ac.uk

Abstract

We propose an unsupervised system that learns continuous degrees of lexicality for noun-noun compounds, beating a strong baseline on several tasks. We demonstrate that the distributional representations of compounds and their parts can be used to learn a fine-grained representation of semantic contribution. Finally, we argue such a representation captures compositionality better than the current status-quo which treats compositionality as a binary classification problem.

1 Introduction

A Multiword Expressions (MWE) can be defined as a sequence of words whose meaning cannot necessarily be derived from the meaning of the words making up that sequence, for example:

Rat Race — self-defeating or pointless pursuit[1]

MWEs are considered a "key problem for the development of large-scale, linguistically sound natural language processing technology" (Sag et al., 2002). The challenge posed by MWEs is threefold, consisting of MWE identification, classification and interpretation. Following the identification of a MWE, it needs to be established whether the expression should be treated as lexical (idiomatic) or as compositional. The final step, learning the semantics of the MWE, strongly depends on this decision.

[1]Definition taken from Wikipedia, and clearly not recoverable if one only knows the meaning of the words 'rat' and 'race'.

The problem posed by MWEs is considered hard, but at the same time it is highly relevant and interesting. MWEs occur frequently in language and interpreting them correctly would directly improve results in a number of tasks in NLP such as translation and parsing (Korkontzelos and Manandhar, 2010). By extension this makes deciding the lexicality of MWEs an important challenge for various fields including machine translation, question answering and information retrieval. In this paper we discuss compositionality with respect to noun-noun compounds.

Most Computational Linguistics literature treats compositionality as a binary problem, classifying compounds as either lexical or compositional. We show that this approach is too simplistic and argue for the real-valued treatment of compositionality.

We propose two unsupervised models that learn compositionality rankings for compounds, placing them on a scale between lexical and compositional extremes. We develop a fine-grained representation of compositionality using a novel generative approach that models context as generated by compound constituents. This representation differentiates between the semantic contribution of both compound constituents as well as the compound itself.

Comparing it with existing work in the field, we demonstrate the competitiveness of our approach. We evaluate on an existing corpus of noun compounds with ranked compositionality data, as well as on a large corpus with a binary annotation for lexical and compositional compounds. We analyse the impact of data sparsity and propose an interpolation approximation which significantly reduces the effect of sparsity on model performance.

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 132–141,
Montréal, Canada, June 7-8, 2012.

2 Related Work

Interpreting MWEs is a difficult task as "compound nouns can be freely constructed" (Spärck Jones, 1985), and are thus able to proliferate infinitely. At the same time, semantic composition can take many different forms, making uniform interpretation of compounds impossible (Zanzotto et al., 2010).

Most current work on MWEs focuses on interpreting compounds and sidesteps the task of determining whether a compound is compositional in the first place (Butnariu et al., 2010; Kim and Baldwin, 2008). Such methods, aimed at learning the semantics of compounds, can roughly be divided into two major strands of research.

One group relies on data intensive methods to extract semantics vectors from large corpora (Baroni and Zamparelli, 2010; Zanzotto et al., 2010; Giesbrecht, 2009). The focus of these approaches is to develop methods for composing the vectors of unigrams into a semantic vector representing a compound. Some of the work in this area touches on the issue of lexicality, as models learning distributional representations of MWEs ideally would first establish whether a given MWE is compositional or not (Mitchell and Lapata, 2010).

The other group are knowledge intensive approaches collecting linguistic features (Kim and Baldwin, 2005; Korkontzelos and Manandhar, 2009). Tratz and Hovy (2010), for instance, train a classifier for noun compound interpretation on a large set of WORDNET and Thesaurus features.

Combined approaches include Kim and Baldwin (2008), who interpret noun compounds by extrapolating their semantics from observations where the two nouns forming a compound are in an intransitive relationship. For example extracting the phrase 'the family owns a car' from the training data would help learn that the compound 'family car' describes a POSSESSOR-OWNED/POSSESSED relationship.

Some of these supervised classifiers include lexicality as a classification option, considering it jointly with the actual compound interpretation.

Next to the work on MWE interpretation there has been some work focused on determining lexicality in its own right (Reddy et al., 2011; Bu et al., 2010; Kim and Baldwin, 2007).

One possibility is to exploit special properties of lexical MWEs such as high statistical association of their constituents (Pedersen, 2011) or syntactic rigidity (Fazly et al., 2009; McCarthy et al., 2007). However, these approaches are limited in their applicability to compound nouns (Reddy et al., 2011).

Another method is to compare the semantics of a compound and its constituents to decide compositionality. The approaches used to determine those semantics can again be divided into knowledge intensive and data-driven methods. Depending on the chosen representation of semantics these approaches can either be used for supervised classifiers or together with a distance metric comparing vector space representations of semantics. In a binary setting, a threshold would then be applied to the result of that distance function (Korkontzelos and Manandhar, 2009). In a real-valued setting the distance metric itself can be used as a measure for compositionality (Reddy et al., 2011). Related to the vector space based models, some research focuses on improving the distance metrics used to compare induced semantics (Bu et al., 2010).

3 Methodology

English noun-noun compounds are majority left-branching (Lauer, 1995), with a head (the second element), modified by an attributive noun (first element). For example:

Ground Floor — The floor of a building at or nearest ground level.[2]

In this paper, we will use the terms attributive noun (AN) and head noun (HN) to refer to the first and second noun in a noun compound.

3.1 Real-Valued Representation

Lexicality of MWEs is frequently treated as a binary property (Tratz and Hovy, 2010; Ó Séaghdha, 2007). We argue that lexicality should instead be treated as a graded property, as most compound semantics exhibit a mixture of compositional and lexical influences. For example, '*cocktail dress*' derives a large part of its semantics from '*dress*', but the compound also contributes an idiosyncratic element to its meaning.

[2]Definition from http://www.thefreedictionary.com

We define lexicality as the degree to which idiosyncrasy contributes to a compound's semantics. Inversely phrased, the compositionality of a compound can be defined as the degree to which its sense is related to the senses of its constituents.[3]

This graded representation follows Spärck Jones (1985), who argued that "it is not possible to maintain a principled distinction between lexicalised and non-lexicalised compounds". Some recent work also supports this view (Reddy et al., 2011; Bu et al., 2010; Baldwin, 2006). From a practical perspective, a real-valued representation of compositionality should help improve interpretation of compounds. This is especially true when factoring in the respective semantic contributions of its parts.

3.2 Context Generation

According to the distributional hypothesis, the semantics of a lexical item can be expressed by its context. We apply this hypothesis to the problem of noun compound compositionality by using a generative model on compound context. Our model allows context to be generated by the compound itself or by either one of its constituents. By learning which element of the compound generates which part of its context we effectively determine the semantic contribution of each element. This in turn gives us a fine-grained, graded representation of a compound's lexicality.

4 Corpora for Evaluation

4.1 Ranked Corpus — REDDY

As we want to evaluate our models' ability to learn lexicality as a real-valued property, we require an annotated data set of noun compounds ranked by lexicality. To the best of our knowledge the only such data set was developed by Reddy et al. (2011). This data set contains 90 distinct noun compounds with real-valued gold standard scores ranking from 0 (lexical) to 5 (compositional). The compounds are nearly linearly distributed across the [0;5] range, with inter annotator agreement (Spearman's ρ) of 0.522. We refer to this data set and evaluation as REDDY throughout this paper.

4.2 Binary Corpora — TRATZ

We also apply our models to a second, binary classification task. Tratz and Hovy (2010) compiled a data set for noun compound interpretation, which classifies noun compounds based on their internal structure. We use this corpus to extract lexical and compositional noun compounds.

After some pre-processing[4] the data set contains 18,858 compositional and 118 lexical noun compounds. We believe this to more accurately represent the real world distribution of lexical and compositional noun compounds: Tratz and Hovy (2010) extracted noun compounds from several large corpora including the Wall Street Journal section of the Penn Treebank, thus obtaining a reasonable approximation of real world occurrence. Other collections of noun compounds (Ó Séaghdha, 2007) feature similar proportions of lexical and compositional noun compounds.

The large bias towards compositional noun compounds does not support the status-quo of treating compositionality as a binary property. As discussed earlier, we assume that most compounds have a compositional as well as a lexical element. While the compositional aspect may be larger for most compounds this alone does not suffice as a reason to disregard the lexical element contained in these compounds.

In order to evaluate our system on the TRATZ data, we use receiving operator characteristic (ROC) curves. ROC analysis enables us to evaluate a ranking model without setting an artificial threshold for the compositionality/lexicality decision.

5 Baseline Approach

We develop a set of advanced baselines related to the semi-supervised models presented by Reddy et al. (2011). We define the context K of a noun compound as all words in all sentences the compound appears in. From this we calculate distributional representations of a compound ($c = \langle a, h \rangle$) and its constituent elements a, h. We refer to these representations as $\vec{c}$ for the compound and $\vec{a}$, $\vec{h}$ for the

[3] For example, the meaning of '*gravy train*' has hardly any relation to either '*gravy*' or '*train*'. Its semantics are thus highly dependent on the compound in its own right. On the other end of the spectrum, '*climate change*' is significantly related to both '*climate*' and '*change*', contributing little inherent semantics to its overall meaning.

[4] We removed trigrams from the data set.

Name	$\oplus$	r	ρ
ADD	$w.S_{ac} + (1-w).S_{hc}$	.323	**.567**
MULT	$S_{ac}.S_{hc}$	**.379**	.551
MIN	$min(S_{ac}, S_{hc})$	.343	.550
MAX	$max(S_{ac}, S_{hc})$	.299	.505
COMB	$w_1.S_{ac}+w_2.S_{hc}+w_3.S_{ac}.S_{hc}$	.366	.556

Table 1: Results of COSLEX with different operators on the REDDY data set, reporting Pearson's r and Spearman's ρ correlations. Weights for operators ADD ($w = 0.3$) and COMB ($\mathbf{w} = \langle 0.3, 0.1, 0.6 \rangle$) are manually optimised. Values range from -1 (negative correlation) to +1 (perfect correlation) with 0 describing random data.

attributive and head noun, respectively. We can calculate the cosine similarity based lexicality score (COSLEX) by combining the cosine similarity of the compound's distribution with each of its two constituents (Reddy et al., 2011).

$$S_{ac} = \text{sim}(\vec{a}, \vec{c})$$
$$S_{hc} = \text{sim}(\vec{h}, \vec{c})$$
$$\text{COSLEX}(c) = S_{ac} \oplus S_{hc}$$

We evaluate a number of alternative operators $\oplus$ for combining S_{ac} and S_{hc}. Results for this baseline on the REDDY corpus are in Table 1,[5] with weights w_i on the combination operators manually optimised for Spearman's ρ on that data set. In effect this renders this baseline into a supervised approach, so we would expect it to perform very well. We use the best performing operators (ADD with $w = 0.3$, MULT) as baselines for this paper.

6 Generative Models

We exploit the distributional hypothesis to model the semantic contribution of the different elements of a noun compound. For this, we require a system that treats a noun compound as a vector of three semantics-bearing units: the compound itself, its head and its attributive noun. This system should then model the relationship between the context of the compound and these three units, deciding which of them is responsible for each context element.

[5]Reddy et al. (2011) report higher figures on our baseline models. The differences are attributed to differences in training data and parametrization.

6.1 3-way Compound Mixture

We model a corpus $\mathcal{D}$ of tuples $d = \{c, k_1, ..., k_n\}$. Each tuple d contains a noun compound $c = \langle a, h \rangle$ and its context words $\mathbf{K} = (k_1, ..., k_n)$. We use vocabularies V_c for noun compounds, V_a for attributive nouns, V_h for head nouns and V_k for context.

We condition our generative model on the noun compounds. Given an observation d of a compound c, we generate each context word in two steps. First, we choose one of the compounds three elements[6] to generate the next context word. Second, we generate a new context word conditioned on that element. Formally, the context is generated as follows.

We draw three multinomial parameters Ψ^c, Ψ^a and Ψ^h from Dirichlet distributions with parameters α^c, α^a and α^h. Ψ^c represents the distribution over context words V_k given compound c. Ψ^a and Ψ^h are distributions over V_k given attributive noun a and head noun h, respectively. These three distributions form the mixture components of our model.

A fourth multinomial parameter Ψ^z, drawn from a Dirichlet distribution with parameter α^z, controls the distribution over the mixture components. Ψ^z is specific to each compound c, so multiple observations of the same compound share this parameter.

For each context word we draw a mixture component $z_{c,i} \in \{\check{c}, \check{a}, \check{h}\}$ from the multinomial distribution with parameter Ψ^z. $z_{c,i}$ determines which distribution the context word itself will be drawn from. Finally, we draw the context word:

$$\forall i\text{: } k_i \mid \Psi^{\{z_{c,i}\}} \sim \text{Multi}(\Psi^{\{z_{c,i}\}})$$

Thus, for each observation of a compound noun we have a vector $\mathbf{z_c} = \langle \mathbf{z_1}, ..., \mathbf{z_n} \rangle$ detailing how its context words were created either by the compound itself or by one of its constituents. To determine lexicality, we are interested in learning the multinomial parameter Ψ^z, which describes to what extent the compound and its constituents contribute to the generation of the context (i.e. semantics). We can approximate Ψ^z from the vector $\mathbf{z_c}$.

We define the lexicality score $Lex(c)$ for a compound as the percentage of context words created by

[6]The compound itself, its attributive noun and its head noun

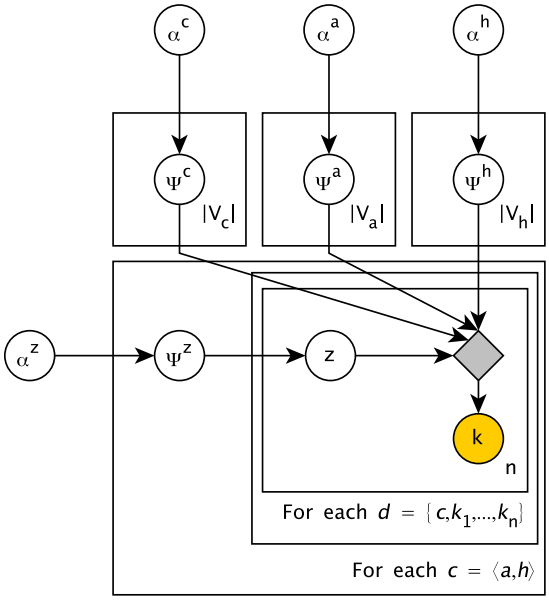

Figure 1: Plate diagram illustrating the MULT-CMPD model with context words k_i drawn from a mixture model with three components controlled by z_i.

the compound and not one of its constituents:

$$Lex(c) = p(z{=}\check{c}|\langle a, h\rangle), \quad \text{where } c = \langle a, h\rangle \tag{1}$$

Figure 1 shows a plate diagram of this model, which we will refer to as MULT-CMPD.

One hypothesis encoded in model MULT-CMPD is that deciding which part of a compound (the compound itself, the head or the attributive noun) generates context is a single decision. An alternative representation could treat this as a two-step process, which we encode in a second model BIN-CMPD. The intuition behind the BIN-CMPD model is that there are two distinct decisions. First, whether a compound is compositional or not. Second, whether (in the compositional case) its semantics stem from its head or attributive noun

Where MULT-CMPD uses a three component mixture to determine which multinomial distribution to use, BIN-CMPD uses two cascaded binary mixtures (see Figure 2). The BIN-CMPD model first chooses whether to treat a compound as compositional or lexical. If the compound is determined as compositional, a second binary mixture determines whether to generate a context word using the attributive (Ψ^a) or head multinomial (Ψ^h). For the lexical case, the model remains unchanged.

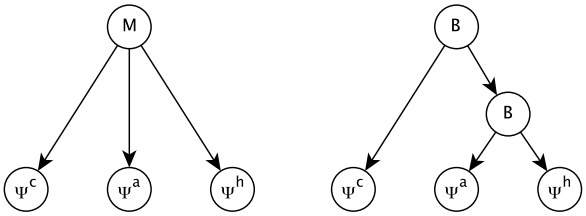

Figure 2: Schematic description of compositionality/lexicality decision for models MULT-CMPD and BIN-CMPD.

Model	r	ρ
COSLEX (ADD)	.323	**.567**
COSLEX (MULT)	**.379**	.551
MULT-CMPD	.141	.435
BIN-CMPD	.168	.410

Table 2: Results on the REDDY data set, reporting Pearson's r and Spearman's ρ correlations. Values range from -1 (negative correlation) to +1 (perfect correlation).

6.1.1 Inference and Sampling

We use Gibbs sampling to learn the vectors **z** for each instance d, integrating out the parameters Ψ^x. We train our models on the British National Corpus (BNC), extracting all noun-noun compounds from a parsed version of the corpus.

In order to speed up convergence of the sampler, we use simulated annealing over the first 20 iterations (Kirkpatrick et al., 1983), helping the randomly initialised model reach a mode faster. We report results using marginal distributions after a further 130 iterations, excluding the counts of the annealing stage.

6.1.2 Evaluation

We evaluate our two models on the REDDY data set by comparing its scores for lexicality ($Lex(c)$) with the annotated gold standard. The aim of this evaluation is to determine how accurately the models can capture gradual distinctions in lexicality. The ROC analysis on the TRATZ data set furthermore informs us how precise the models are at distinguishing lexical from compositional compounds.

Results of the REDDY evaluation are in Table 2. We use Spearman's ρ to measure the monotonic correlation of our data to the gold standard. Pearson's r additionally captures the linear relationship between the data, taking into account the relative differences in $Lex(c)$ scores among noun compounds.

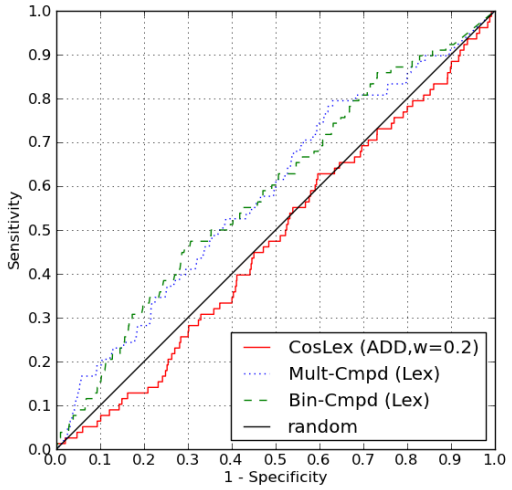

Figure 3: ROC analysis of models MULT-CMPD and BIN-CMPD versus the best COSLEX baseline (ADD) on the TRATZ data set

While both models, BIN-CMPD and MULT-CMPD, clearly learn a correlation with lexicality rankings, they underperform the strong, semi-supervised COSLEX baselines described earlier in this paper. The second evaluation, on the binary TRATZ data set shows a different picture (see Figure 3). The best COSLEX baseline (ADD with $w = 0.2$) fails to outperform random choice on this task. Both generative models clearly beat COSLEX on this task, with MULT-CMPD in particular performing very well for low sensitivity.

There is no clear distinction in performance between the two generative approaches. Further analysis might help us to separate the two more clearly, and we will continue using both models throughout this paper.

It is important to note the different performance of the generative models vs. the cosine similarity approach on two tasks. The REDDY data set has a nearly linear distribution of compositionality scores, while the TRATZ data set is overwhelmingly compositional, which more closely represents the real world distribution of compounds. The poor performance of the cosine similarity approach (COSLEX) on the TRATZ evaluation suggests the limitations of this approach when applied to more realistic data such as this data set. An additional explanation for the semi-supervised baseline's poorer result is that the effect of parameter tuning decreases on larger data.

Investigating the errors made by the models MULT-CMPD and BIN-CMPD gives rise to a number of possible explanations for their performance. The most promising lead is related to data sparsity, with many of the evaluated noun-noun compounds only appearing once or twice in the corpus. This makes it harder for our generative approach to learn sensible context distributions for these instances.

We will next investigate how to reduce the effects encountered by sparsity.

6.2 Interpolation

Working on problems related to non-unigram data, sparsity is a frequently encountered problem. As already explored in the previous section, this is also the case for our generative models of lexicality.

It would be possible to use an even larger training corpus, but there are limitations as to what extent this is possible. The BNC, containing 100 million words, is already one of the largest corpora regularly used in Computational Linguistics. However, adding more data in an unsupervised sense is unlikely to significantly improve results (Brants et al., 2007).

Alternatively, it would be possible to add specific training data that included the noun compounds from the evaluation data sets. This would, however, compromise the unsupervised nature of our approach, and it thus not an option either.

In this paper, we will instead focus on extenuating the effects of data sparsity through other unsupervised means. For this purpose we investigate interpolating on a larger set of noun compounds.

Kim and Baldwin (2007) observed that semantic similarity of verb-particle compounds correlates with their lexicality. We extend this observation for noun compounds, hypothesising that the lexicality of similar words will be similar. We combine this with the assumption that noun compounds sharing a constituent are likely to be semantically similar (Korkontzelos and Manandhar, 2009).

Using this idea, we can approximate the lexicality of a given compound with the lexicality scores of all compounds sharing either of its constituents. So far we have calculated the lexicality of a given compound using the formula $Lex(c)$ in Equation 1. The formula $Clex(c)$ in Equation 2 averages the lexicality scores of a compound with those of its related

Function and Model		r	ρ
	CosLex (ADD)	.323	.567
	CosLex (MULT)	.379	.551
$Lex(c)$	Mult-Cmpd	.141	.435
	Bin-Cmpd	.168	.410
$Clex(c)$	Mult-Cmpd	.357	.596
	Bin-Cmpd	.400	.592
$Ilex(c)$	Mult-Cmpd	.422	.621
	Bin-Cmpd	**.538**	**.623**

Table 3: Results on the REDDY data set, reporting Pearson's r and Spearman's ρ correlations, comparing $Ilex(c)$ and $Clec(c)$ interpolations with $Lex(c)$.

compounds. As $p(z{=}1|\langle a, h\rangle)$ directly influences both $p(z{=}1|\langle a, \cdot\rangle)$ and $p(z{=}1|\langle \cdot, h\rangle)$, we can also consider dropping it from the approximation such as in Equation 3. This approach trades some specificity in favour of reducing sparsity, as we observe more instances of such related compounds than of a particular noun compound itself only.

$$Lex(c) \approx Clex(c) \quad (2)$$
$$Clex(c) = \frac{p(z{=}1|\langle a, \cdot\rangle) + p(z{=}1|\langle \cdot, h\rangle) + p(z{=}1|\langle a, h\rangle)}{3},$$
where $c = \langle a, h\rangle$
$$Lex(c) \approx Ilex(c) \quad (3)$$
$$Ilex(c) = \frac{p(z{=}1|\langle a, \cdot\rangle) + p(z{=}1|\langle \cdot, h\rangle)}{2},$$
where $c = \langle a, h\rangle$

Both formulations enable us to better deal with sparse data as decisions are made based on a wider range of observations. At the same time, we avoid a loss of specificity as the models and scores are still highly dependent on the individual noun compound.

We avoid introducing additional degrees of freedom by using uniform weights only. However, it would be simple to turn this approach into a semi-supervised model by tuning the weights for the different probabilities involved in calculating $Clex(c)$ and $Lex(c)$. That approach would be comparable to the operators used on our CosLex baselines.

Results on the REDDY data set using $Clex(c)$ and $Ilex(c)$ are in Table 3. Figure 4 shows the impact of these approximations on the Tratz data for the Bin-Cmpd model. These interpolations suggest strong improvements in performance. It should especially be noted that $Ilex(c)$ consistently outperforms $Clex(c)$, which indicates the strength of the

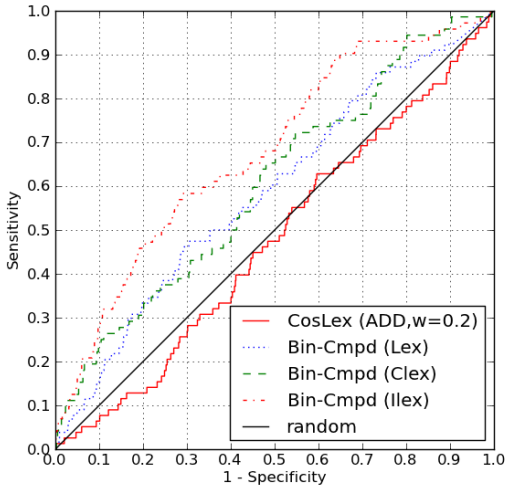

Figure 4: ROC analysis of model Bin-Cmpd on the TRATZ data set, comparing $Ilex(c)$ and $Clec(c)$ interpolations with $Lex(c)$.

related-compound probabilities over the individual compound probabilities.

These results confirm our suspicion that sparsity was a major factor affecting our models' performance. Furthermore, they strengthen our hypothesis about the relatedness of semantic similarity and lexicality and demonstrate a sensible approach for exploiting this relationship.

7 Analysis

We use this section for qualitative evaluation, complementing the quantitative evaluation in the previous sections. The purpose of the qualitative evaluation is to better understand exactly what it is our models are learning.

Table 5 lists the compounds that model Bin-Cmpd considers the most lexical and the most compositional. The list of compounds with the high lexicality scores is dominated by proper nouns such as countries, companies and persons. This is in line with expectation as compounds of proper nouns are fully lexical. Removing proper nouns (also in Table 5), we get a slightly more ambiguous list. For example, 'study design' is not considered a lexical compound, but rather a highly institutionalized, compositional MWE (Sag et al., 2002). Using $Lex(c)$ 'study design' is ranked as such, so this appears to be a case where interpolation has a negative impact.

In this paper we argued for a finer grained analysis of compositionality, taking into account the differ-

Context of 'flea market' generated by

flea	**market**	**flea market**
canal, wall, incline, campsite	stall, Paris, sale, Saturday, week, Sunday, quarter, damage, change	barter, souvenir, launderette, Lamine, Canet, Kouyate, Plage

Context of 'memory lane' generated by

memory	**lane**	**memory lane**
take, story, about, tell, real, glimpse, Britain, reminiscence	village, protection, drive, catwalk, plant	war, justify, bill, Campbell, rude-boys

Context of 'night owl' generated by

night	**owl**	**night owl**
court, fee, guest, early, day, Baden, membership, life, game	waive, player, Halikarnas, bar, bird, unbooked, Vienna	adventurous

Context of 'melting pot' generated by

melting	**pot**	**melting pot**
forest, racial, caribbean, plan, programme, reality, arrangement	in, into, put, political, community, prepare	ethnic, greatest, drawing, liaise, pan-european, myth

Table 4: Overview over context words generated by model BIN-CMPD. We list a selection of words predominately generated by each of the mixture components of the given noun-noun compound.

Most Compositional
labour union, tax authority, health council, market counterparty, employment policy

Most Lexical
study design, family motto, wood shaving, avoidance behaviour, smash hit

Most Lexical (including Proper Nouns)
Vo Quy, Bonito Oliva, Mamur Zapt, Evander Holyfield, Saudi Arabia

Table 5: Top lexical and compositional nouns for the BIN-CMPD model using $Ilex(c)$

ent impact of both constituents. We tried to achieve this by modelling a compound's context as generated from its various semantic constituents. Table 4 highlights the impact of this method for a number of noun compounds, showing which context words were predominately generated by each constituent.

Due to the nature of the context used, some of the links are semantically not obvious (e.g. the relationship between owls and Vienna). In some cases the semantic contribution of the parts is more clearly separated, such as the contributions of '*memory*' and '*lane*' to the semantics of '*memory lane*'. In summary, these examples clearly suggest that our models learn to associate context with compound elements and that this association is an informed one.

8 Conclusion

We proposed a novel approach for learning lexicality scores for noun compounds and empirically demonstrated the feasiblity of this approach. Using a generative model we were able to beat a strong, semi-supervised baseline with an unsupervised model.

We discussed the issue of data sparsity in depth and proposed several approaches for overcoming this problem. Focusing on unsupervised approaches, we demonstrated how interpolation can be used to tackle sparsity. The two interpolation methods that we implemented helped us to strongly improve overall model performance. Our empirical evaluation of interpolation metrics $Clex(c)$ and $Ilex(c)$ also gives credence to the hypothesis that lexicality is related to semantic similarity.

On the theoretical side, we offered further support to the real-valued treatment of lexicality.

Further work will include using larger training corpora. While the BNC is a popular corpus in Computational Linguistics, it proved to be too small to learn sensible representations for a number of compounds encountered in the test data. Using larger corpora will also allow us to further study and reduce the sparsity issues encountered.

To study the relationship between constituent and compound compositionality in greater depth, we will also investigate alternative approaches for interpolation. Similarity measures that consider the semantic relevance of individual context elements should also be considered as a next step.

Another obvious source of future work is to apply our approach to general collocations beyond the special case of noun compounds only.

Acknowledgments

The authors would like to acknowledge the use of the Oxford Supercomputing Centre (OSC) in carrying out this work.

References

Timothy Baldwin. 2006. Compositionality and multiword expressions: Six of one, half a dozen of the other? In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, page 1, Sydney, Australia. Association for Computational Linguistics.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1183–1193, Stroudsburg, PA, USA. Association for Computational Linguistics.

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large Language Models in Machine Translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867.

Fan Bu, Xiaoyan Zhu, and Ming Li. 2010. Measuring the non-compositionality of multiword expressions. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 116–124, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cristina Butnariu, Su Nam Kim, Preslav Nakov, Diarmuid Ó. Séaghdha, Stan Szpakowicz, and Tony Veale. 2010. Semeval-2010 task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 39–44, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

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, pages 173–184, Berlin, Heidelberg. Springer-Verlag.

Su Nam Kim and Timothy Baldwin. 2005. Automatic interpretation of noun compounds using wordnet similarity. In *In Proceedings of the 2nd International Joint Conference on Natural Language Processing, Jeju Island, South Korea, 1113*, pages 945–956.

Su Nam Kim and Timothy Baldwin. 2007. Detecting compositionality of English verb-particle constructions using semantic similarity. In *Proceedings of the 7th Meeting of the Pacific Association for Computational Linguistics*, PACLING '07, pages 40–48.

Su Nam Kim and Timothy Baldwin. 2008. An unsupervised approach to interpreting noun compounds. In *Natural Language Processing and Knowledge Engineering, 2008. NLP-KE '08. International Conference on*, pages 1–7.

S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. 1983. Optimization by simulated annealing. *Science*, 220(4598):671–680.

Ioannis Korkontzelos and Suresh Manandhar. 2009. Detecting compositionality in multi-word expressions. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, pages 65–68, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ioannis Korkontzelos and Suresh Manandhar. 2010. Can recognising multiword expressions improve shallow parsing? In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 636–644, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mark Lauer. 1995. Corpus statistics meet the noun compound: some empirical results. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, ACL '95, pages 47–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

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, Prague, Czech Republic. Association for Computational Linguistics.

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

Diarmuid Ó Séaghdha. 2007. Annotating and learning compound noun semantics. In *Proceedings of the 45th Annual Meeting of the ACL: Student Research Workshop*, ACL '07, pages 73–78, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ted Pedersen. 2011. Identifying collocations to measure compositionality: shared task system description. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, DiSCo '11, pages 33–37, Stroudsburg, PA, USA. Association for Computational Linguistics.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in com-

pound nouns. In *Proceedings of The 5th International Joint Conference on Natural Language Processing 2011 (IJCNLP 2011)*, Chiang Mai, Thailand.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *In Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002*, pages 1–15.

Karen Spärck Jones. 1985. Compound noun interpretation problems. In Frank Fallside and William A. Woods, editors, *Computer speech processing*, pages 363–381. Prentice Hall International (UK) Ltd., Hertfordshire, UK, UK.

Stephen Tratz and Eduard Hovy. 2010. A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 678–687, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fabio Massimo Zanzotto, Ioannis Korkontzelos, Francesca Fallucchi, and Suresh Manandhar. 2010. Estimating linear models for compositional distributional semantics. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1263–1271, Stroudsburg, PA, USA. Association for Computational Linguistics.