Learning Compositionality Functions on Word Embeddings for Modelling Attribute Meaning in Adjective-Noun Phrases

Matthias Hartung, Fabian Kaupmann, Soufian Jebbara, and Philipp Cimiano

Semantic Computing Group
CITEC, Bielefeld University

{mhartung,fkaupmann,sjebbara,cimiano}@techfak.uni-bielefeld.de

Abstract

Word embeddings have been shown to be highly effective in a variety of lexical semantic tasks. They tend to capture meaningful relational similarities between individual words, at the expense of lacking the capabilty of making the underlying semantic relation explicit. In this paper, we investigate the *attribute* relation that often holds between the constituents of adjective-noun phrases. We use CBOW word embeddings to represent word meaning and learn a compositionality function that combines the individual constituents into a phrase representation, thus capturing the compositional attribute meaning. The resulting embedding model, while being fully interpretable, outperforms count-based distributional vector space models that are tailored to attribute meaning in the two tasks of attribute selection and phrase similarity prediction. Moreover, as the model captures a generalized layer of attribute meaning, it bears the potential to be used for predictions over various attribute inventories without re-training.

1 Introduction

Attributes such as SIZE, WEIGHT or COLOR are part of the building blocks of representing knowledge about real-world entities or events (Barsalou, 1992). In natural language, formal attributes find their counterpart in attribute nouns which can be used in order to generalize over individual properties, e.g., *big* or *small* in case of SIZE, *blue* or *red* in case of COLOR (Hartung, 2015).

In order to ascribe such properties to entities or events, adjective-noun phrases are a very frequent linguistic pattern. In these constructions, attribute meaning is conveyed only implicitly, i.e., without being overtly realized at the phrasal surface. Hence, *attribute selection* has been defined as the task of predicting the hidden attribute meaning expressed by a property-denoting adjective in composition with a noun (Hartung and Frank, 2011b), as in the following examples:

(1) a. *hot summer* → TEMPERATURE
b. *hot debate* → EMOTIONALITY
c. *hot soup* → TASTE/TEMPERATURE

Previous work on this task has largely been carried out in distributional semantic models (cf. Hartung (2015) for an overview). In the face of the recent rise of distributed neural representations as a means of capturing lexical meaning in NLP tasks (Collobert et al., 2011; Mikolov et al., 2013a; Pennington et al., 2014), our goal in this paper is to model attribute meaning based on word embeddings. In particular, we use CBOW embeddings of adjectives and nouns (Mikolov et al., 2013a) as underlying word representations and train a compositionality function in order to compute a phrase representation that is predictive of the implicitly conveyed attribute meaning.

In fact, word embeddings (also referred to as *predict models*) have been shown to be highly effective in a variety of lexical semantic tasks (Baroni et al., 2014b), compared to "traditional" distributional semantic models (or *count models*) in the tradition of Harris (1954). However, this finding has been refuted to a certain extent by Levy et al. (2015), stating that much of the perceived superiority of word embeddings is due to hyperparameter optimizations rather than principled advantages. Moreover, the authors found that in many cases, tailoring count models to a particular task at hand is both feasible and beneficial in order to outperform the more generic embeddings.

This sheds light on a definitive plus of count models, viz. their transparency and interpretability in the sense that their semantic similarity ratings can (under certain conditions) be traced back to particular semantic relations, whereas word embeddings typically yield rather vague and diversified similarities (Erk, 2016). Due to this lack in interpretability, word embeddings are not easily interoperable with symbolic lexical resources or ontologies. Thus, we argue that modelling attribute meaning poses an interesting challenge to word embeddings for two reasons: First, being rooted in ontological knowledge, attribute meaning clearly draws on interpretability of the underlying model; second, attribute meaning in adjective-noun phrases is conveyed in compositional processes (cf. Ex. (1)) which are underresearched in the context of word embeddings so far (Manning, 2015).

Our main contributions in this paper are: (i) We demonstrate that word embeddings can be successfully harnessed for attribute selection – a task that requires both compositional and interpretable representations of phrase meaning. (ii) This is achieved via a learned compositionality function f on adjective and noun embeddings that carves out attribute meaning in their compositional phrase meaning. (iii) We show that f captures generalized attribute meaning (cf. Bride et al. (2015)) that abstracts from individual attributes. Thus, after fitting the compositionality function, our model bears the potential of being applied to various application scenarios (e.g., aspect-based sentiment analysis) involving diverse attribute inventories. (iv) We show that the same model also scales to the task of predicting semantic similarity of adjective-noun phrases, which indicates both the robustness of the model and the importance of attribute meaning as a major source of phrase similarity.

2 Related Work

Attribute Learning from Adjectives and Nouns. Adjective-centric approaches to attribute learning from text date back to Almuhareb (2006) and Cimiano (2006). Bakhshandeh and Allen (2015) present a sequence tagging model in order to extract attribute nouns from adjective glosses in WordNet. Most recently, Petersen and Hellwig (2016) use a clustering approach based on adjective-noun co-occurrences in order to induce clusters of German adjectives that constitute the value space of an attribute. However, their approach falls short of making the respective attribute explicit.

These approaches have in common that they do not consider the compositional semantics of an adjective in its phrasal context with a noun in order to derive attribute meaning. This is in contrast to Hartung and Frank (2010; 2011b) who frame attribute selection in a distributional count model which (i) encodes adjectives and nouns as distributional word vectors over attributes as shared dimensions of meaning and (ii) uses vector mixture operations in order to compose these word vectors into phrase reresentations that are predictive of compositional attribute meaning.

Tandon et al. (2014) propose a semi-supervised method for populating a knowledge base with triples of nouns, attributes and adjectives that are acquired from adjective-noun phrases. Being based on label propagation over monosemous adjectives as seeds, their approach depends on a lexical resource providing initial mappings between adjectives and attributes.

The present approach and the work by Hartung and Frank may be considered as pairs of opposites in two respects: First, our model is based on pretrained CBOW word embeddings for representing adjective and noun meaning. Thus, we do not encode any attribute-specific lexical information explicitly at the level of word representation. Second, we apply function learning in order to empirically induce a compositionality function that is trained to promote aspects of attribute meaning in adjective-noun phrase embeddings.

Compositionality. Modelling compositional processes at the intersection of word and phrase meaning in distributional semantic models has attracted considerable attention in the last years (Erk, 2012). Mitchell and Lapata (2010) have promoted a variety of vector mixture models for the task, which have been criticized for their syntactic agnosticism (Baroni and Zamparelli, 2010; Guevara, 2010).

Focussing on adjective-noun compositionality, the latter authors propose instead to model adjective meaning as matrices encoding linear mappings between noun vectors. These attempts to integrate formal semantic principles in the tradition of Frege (1892) into a distributional framework have been generalized to a "program for compositional distributional semantics" (Baroni et al.,

2014a) that is centered around *functional application* as the general process to model compositionality in semantic spaces, thus emphasizing the insight that different linguistic phenomena require to be modeled in corresponding algebraic structures and composition operators matching these structures (cf. Widdows (2008), Grefenstette and Sadrzadeh (2011), Grefenstette et al. (2014)).

Bride et al. (2015) observe that such composition operators, by being trained on empirical corpus data, can either be tailored to specific lexical types (i.e., individual composition functions for each adjective in the corpus), or designed to capture general compositional processes in syntactic configurations (i.e., a single lexical function for all adjective-noun phrases). In line with these authors, we aim at learning a lexical function which captures attribute meaning in the compositional semantics of adjective-noun phrases, while *generalizing* over individual attributes.

Contrary to distributional count models, there is relatively few work on applying word embeddings to linguistic problems or NLP tasks related to compositionality. Notable exceptions are Socher et al. (2013) for sentiment analysis, as well as Salehi et al. (2015) and Cordeiro et al. (2016) who focus on predicting the degree of compositionality in nominal compounds rather than carving out a particular semantic relation that is expressed in their compositional semantics.

3 Learning Attribute Meaning in Word Embeddings

3.1 Attribute Meaning in Natural Language

Natural language refers to ontological attributes in terms of attribute nouns such as *color*, *size* or *shape* (Guarino, 1992; Löbner, 2013). Therefore, despite remaining mostly implicit in adjective-noun phrases (cf. Ex. (1) above), we hypothesize that attribute meaning can be learned from contextual patterns of attribute nouns in natural language text. This leads us to the assumption that *adjectives, nouns and attributes (via attribute nouns) can be embedded in the same semantic space.*

3.2 Compositional Models of Attribute Meaning

In this work, we aim at a compositional approach to attribute meaning in adjective-noun phrases. As a consequence of the above assumption, our model represents adjectives, nouns and attributes as vectors $\vec{a}$, $\vec{n}$ and $\vec{attr}$, respectively, in one and the same embedding space $\mathcal{S} \subseteq \mathbb{R}^d$.

By designing a composition function $f(\vec{a}, \vec{n})$ that produces phrase representations $\vec{p} \in \mathcal{S}$, we can use nearest neighbour search in $\mathcal{S}$ in order to predict the attribute $\widehat{attr}$ that is most likely expressed in the compositional semantics of an adjective-noun phrase p:

$$\widehat{attr} := \arg \max_{attr \in A} \cos(\vec{p}, \vec{attr}) \qquad (2)$$

where $\vec{p} = f(\vec{a}, \vec{n})$, cos denotes cosine vector similarity and A the set of all attributes considered. The compositional functions that we use in this work can be divided into baseline models, largely derived from Mitchell and Lapata (2010), and trainable models.

3.2.1 Baseline Models

Adjective or Noun. The simplest model is to skip any composition and just use the representation of the adjective or the noun as a surrogate: $\vec{p} = \vec{a}$ or $\vec{p} = \vec{n}$, respectively.

Pointwise Vector Addition. The first step in the direction of compositionality is pointwise vector addition: $\vec{p} = \vec{a} + \vec{n}$. According to Mitchell and Lapata (2010), the commutativity of addition is a disadvantage because the model ignores word order and thus syntactic information is lost.

Weighted Vector Addition. For the latter reason, Mitchell and Lapata (2010) also propose a weighted variant of pointwise vector addition. In order to account for possibly different contributions of the constituents to phrasal composition, scalar weights α and β are applied to the word vectors before pointwise addition: $\vec{p} = \alpha\vec{a} + \beta\vec{n}$.

Pointwise Vector Multiplication. This composition function multiplies the individual dimensions of the adjective and noun vector: $p_i = a_i \cdot b_i$. Mitchell and Lapata (2010) point out that vector multiplication can be seen as equivalent to logical intersection. In previous work on attribute selection in a count-based distributional framework, the best results were obtained using pointwise multiplication (Hartung, 2015).

Dilation. The dilation model of Mitchell and Lapata (2010) dilates one vector in the direction of the other. This is inspired by the dilation effect of matrix multiplication, but is specifically designed

to be basis-independent:

$$\vec{p} = (\vec{n} \cdot \vec{n})\vec{a} + (\lambda - 1)(\vec{n} \cdot \vec{a})\vec{a} \quad (3)$$

Here, $\vec{n}$ is stretched by a factor λ to emphasize the contribution of $\vec{a}$. λ is a parameter that has to be chosen manually. Analogously, dilation of the adjective is possible as well.

3.2.2 Trainable Models

In this section, we present a method for supervised training of compositionality functions. We propose additive and multiplicative models that use weighting matrices or tensors to balance the contributions of adjectives and nouns. The composition is trained to specifically capture attribute meaning in the resulting phrase representation. The weights are trained as part of a shallow neural network (see Section 3.2.3).

Full Weighted Additive Model. Following Guevara (2010), the full additive model capitalizes on vector addition with weighting matrices for adjective and noun:

$$\vec{p} = \mathbf{A} \cdot \vec{a} + \mathbf{N} \cdot \vec{n} \quad (4)$$

As initializations of the weighting matrices, we use an identity matrix[1], which is equivalent to non-parametric vector addition. As weighting schemes, we use one of (i) weighting only the adjective or noun, respectively, or (ii) weighting both adjective and noun distinctly.

Note that, in line with Guevara (2010), this model makes use of weight matrices in order to balance the contribution of adjectives and nouns to phrasal attribute meaning, whereas Mitchell and Lapata (2010) use scalar weights in their pointwise additive model (cf. Section 3.2.1). Our intuition is that full additive models should be better suited to model compositonal processes that involve interactions between dimensions of meaning.

Trained Tensor Product. As a weighted multiplicative model, we use multiplication of adjective and noun representations with a learned third-order tensor $\boldsymbol{T}$, following Bride et al. (2015):

$$\vec{p} = \vec{a}^T \cdot \boldsymbol{T}^{[1:d]} \cdot \vec{n} \quad (5)$$

with $\vec{a} \in \mathbb{R}^d,\ \vec{n} \in \mathbb{R}^d,\ \boldsymbol{T}^{[1:d]} \in \mathbb{R}^{d \times d \times d}$

In order to compose a phrase representation $\vec{p}$ from $\vec{a}$ and $\vec{n}$, $\boldsymbol{T}$ is applied to the adjective vector in a tensor dot product. The tensor dot product multiplies components of vector and tensor and sums along the third axis of the tensor:

$$\mathbf{X}_{i,j} = \sum_{k=1}^{d} a_k \cdot \boldsymbol{T}_{i,j,k} \quad (6)$$

with d being the dimensionality of the word embeddings. Equation (6) results in a matrix $\mathbf{X}$ that is multiplied with the noun vector in a second step using common matrix multiplication: $\vec{p} = \mathbf{X} \cdot \vec{n}$.

Note that the latter step corresponds to *functional application* of the adjective to the noun as rooted in compositional distributional semantics (Baroni et al., 2014a). The result is a phrase vector with the same dimensionality as adjective and noun. For initialization, we use an identity matrix for each second-order tensor along the third axis[2].

3.2.3 Training Method

The weights of the models in Section 3.2.2 are trained as part of a shallow neural network with no hidden layer. For each adjective-noun phrase and the corresponding ground truth attribute in the training dataset, the respective 300-dimensional vectors[3] $\vec{a}$, $\vec{n}$ and $\vec{attr}$ are obtained by performing a look-up in the pre-trained word embeddings.

With $\vec{a}$ and $\vec{n}$ as its inputs, the neural network computes a phrase representation $\vec{p} \in \mathbb{R}^{300}$ at the output layer. The error of the computed phrase representation to the expected attribute representation $\vec{attr}$ is computed using the *mean squared error* between the two vectors and is used as the training signal for the network parameters. Note that we do not train the embedding vectors along with the connection weights. While this could potentially benefit the results, we aim to explore whether generally trained word embeddings can be used to retrieve attribute meaning.

For our network architectures and computations, we use the deep learning library *keras* (Chollet, 2016). Training takes 10 iterations over the training data; weights are optimized using the stochastic optimization method *Adam* (Kingma and Ba, 2015). For the use of pre-trained word

[1] We also experimented with different initializations such as random values, all-ones, or an identity matrix with additional small random values on non-diagonal elements, but found the identity matrix to work best.

[2] We found a random initialization of all entries to perform substantially worse.

[3] This is the number of dimensions in the pre-trained word embeddings from Mikolov et al. (2013b).

vectors (Mikolov et al., 2013b)[4] in a Python environment, we rely on the *Gensim* library (Řehůřek and Sojka, 2010).

4 Attribute Selection Experiments

In this experiment, we evaluate the compositional models defined in Section 3.2 on the attribute selection task.

4.1 Data

We use the HeiPLAS data set (Hartung, 2015) which contains adjective-attribute-noun triples that were heuristically extracted from WordNet (Miller and Fellbaum, 1998) and manually filtered by linguistic curators. The data is separated into development and test set (comprising 869 and 729 triples, respectively, which correspond to a total of 254 target attributes). The target attributes are subdivided into various semantically homogeneous subsets, as shown in Table 1. Due to coverage issues in the pre-trained word2vec embeddings (Mikolov et al., 2013a), some adjectives and nouns from HeiPLAS cannot be projected into the embedding space[5].

4.2 Experiment 1: Large-scale Attribute Selection

Experimental Procedure. Composition models as described in Section 3.2.2 are trained on all triples in HeiPLAS-Dev (following the procedure described in Section 3.2.3) and evaluated on HeiPLAS-Test. The word vector representations corresponding to the adjective and the noun in a test triple are composed into a phrase vector by applying the trained composition function. Using nearest neighbour search in $\mathcal{S}$ as described in Section 3.2, all test attributes are ranked wrt. their similarity to the composed phrase vector. For evaluation, we use *precision-at-rank* to measure the number of times the correct attribute is ranked as most similar to the phrase vector or among the first five ranks (P@1 and P@5, respectively).

Baseline Semantic Spaces. We directly compare our approach against the results of two count-based distributional models, C-LDA and L-LDA (Hartung, 2015), on the same evaluation data. C-LDA and L-LDA induce distributional adjective and noun vectors over attributes as dimensions of meaning, which are composed into phrase representations using pointwise vector multiplication. Using these models for comparison enables us to assess both the impact of different types of word representations (dense CBOW word embeddings vs. specifically tailored attribute-based distributional word vectors) and different approaches to compositionality (pre-defined vector mixture operations on attribute-specific word representations vs. trained composition functions for promoting generalized attribute meaning in word embeddings).

Results. Results of Experiment 1 are shown in Table 2. The upper part of the table contains the results based on word embeddings (comprising non-parametric, parametric, dilation and trainable composition models); the count-based C-LDA and L-LDA baselines are displayed below.

Focussing on the non-parametric models first, we find that relying on the adjective embedding as a surrogate of a composed representation already outperforms both count models by a wide margin. This indicates a clear advantage of CBOW embeddings over count-based representations for capturing attribute meaning at the word level. However, this holds only for adjectives; noun embeddings in isolation perform much worse.

This is confirmed by the dilation results: Dilating the noun representation into the direction of the adjective performs considerably better than vice versa, while there is no improvement beyond the non-compositional adjective baseline. These findings are in line with Hartung (2015) and Hartung and Frank (2011a) who also observed that adjective representations capture more of the compositional attribute semantics in adjective-noun phrases than noun representations do.

Considering the trained composition models, we find that weighting either the adjective or the noun in a full additive model substantially outperforms the respective non-compositional baseline. The overall best results are obtained by assigning trained weights to both the adjective and the noun embedding (P@1=0.56). This model also outperforms weighted vector addition[6] using scalar weights by great margins.

[4] Available from `https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?usp=sharing`

[5] This affects 54 triples in HeiPLAS-Dev and 44 triples in HeiPLAS-Test, which were removed from the evaluation.

[6] The weighted vector addition scores shown in Table 2 are based on optimized parameters as reported by Mitchell and Lapata (2010): α=0.88 and β=0.12. By shifting the parameters further into the direction of the adjective (i.e., α=0.90; β=0.10), P@1 slightly increases to 0.34.

Subset	Num. Attributes	Num. Train. Triples	Example Phrases
Core	10	72	*silvery hair* (COLOR), *huge wave* (SIZE), *longstanding conflict* (DURATION)
Selected	23	153	*sufficient food* (QUANTITY), *grave decision* (IMPORTANCE), *broad river* (WIDTH)
Measurable	65	261	*heavy load* (WEIGHT), *short hair* (LENGTH), *slow walker* (SPEED)
Property	73	300	*young people* (AGE), *high mountain* (HEIGHT), *straight line* (SHAPE)
All	254	869	*dry paint* (WETNESS), *scentless wisp* (SMELL), *vehement defense* (STRENGTH)

Table 1: Overview of subsets of attributes contained in HeiPLAS data, together with example phrases

	Compositional Model	P@1	P@5
predict models	Adjective	0.33	0.50
	Noun	0.03	0.10
	Vector Addition ($\oplus$)	0.24	0.45
	Weighted Vector Addition	0.33	0.51
	Vector Multiplication ($\odot$)	0.00	0.02
	Adj. Dilation ($\lambda = 2$)	0.06	0.18
	Noun Dilation ($\lambda = 2$)	0.33	0.51
	Full Add. Weighted Noun	0.33	0.54
	Full Add. Weighted Adjective	0.46	0.71
	Full Add. Weighted Adj. and Noun	**0.56**	**0.75**
	Trained Tensor Product ($\otimes$)	0.44	0.57
count	C-LDA (Hartung, 2015)	0.09	n/a
	L-LDA (Hartung, 2015)	0.16	n/a

Table 2: Results of Experiment 1; evaluation on all phrases from HeiPLAS-Test

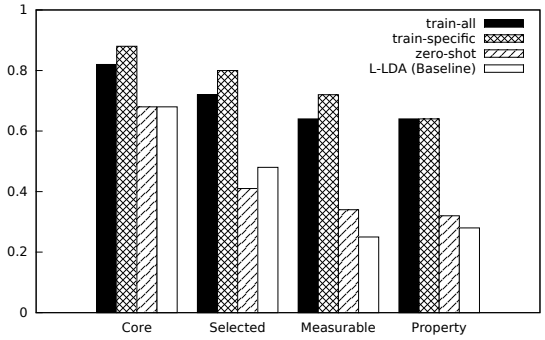

Figure 1: Attribute selection performance of the full additive model after training on all attributes, specific subsets, and in zero-shot learning

In comparison to the best full additive model, the tensor product underperforms by more than 10 points in P@1 and also falls short of weighting only the adjective. This is in line with a general preference of word embeddings for additive models (Mikolov et al., 2013a), which is also confirmed by the non-parametric composition functions. On the other hand, we conjecture that the relatively small size of the training set used here is not sufficient for optimally tuning the 300^3 parameters in the learned tensor.

4.3 Experiment 2: Generalization Power

In this experiment, we are interested in assessing the generalization power of the best-performing composition function as trained in Experiment 1. More precisely, we investigate the hypothesis that a full additive model captures a generalized compositional process in the semantics of attribute-denoting adjective-noun phrases rather than the lexical meaning of individual attributes (cf. Bride et al. (2015)).

We evaluate this hypothesis wrt. (i) the fit of the composition function to different subsets of testing attributes, and (ii) its predictive capacity in a zero-shot learning scenario.

Subsets of Testing Attributes. First, we compare the fit of the composition function that has been trained on all attributes (cf. Experiment 1) on the different subsets of attributes in HeiPLAS-Test, as displayed in Table 1.

The results of this experiment are shown in Figure 1. As can be seen from the solid bars in the plot, the attribute selection performance on individual subsets is considerably stronger than on the entire inventory, ranging from P@1=0.82 on the Core subset to P@1=0.64 on the Property and Measurable subsets (compared to P@1=0.56 on all attributes; cf. Table 2). The cross-hatched bars in the figure indicate the relative differences that result from re-training a composition function on the specific subset of interest. The improvements are consistently small (max. +0.08 on the Selected and Measurable subsets); in case of the Property subset, there is no difference at all.

Zero-Shot Learning. As defined by Palatucci et al. (2009), zero-shot learning is the task of learning a classifier for predicting novel class labels un-

seen during training. In order to assess the selection performance of our model in a zero-shot setting, we create four zero-shot training sets by removing from HeiPLAS-Train all attributes that are contained in each of the subsets described in Table 1, respectively. The corresponding subset from HeiPLAS-Test is used for evaluation afterwards.

The zero-shot results are shown by the diagonally hatched bars in Fig. 1. We find that Core attributes, without being seen during training, can be predicted at a performance of P@1=0.68. On larger subsets, zero-shot performance decreases (down to P@1=0.32 on Property attributes). Yet, we consider these results very decent overall, given that they are largely comparable or even superior (except for the Selected subset) to the best scores of the distributional L-LDA model (Hartung, 2015) as shown by the plain bars in Fig. 1.

Even though benefits from attribute-specific training cannot be denied, we find that the trained compositionality function is largely capable of generalizing over individual target attributes.

4.4 Discussion

Our experiments on attribute selection show that CBOW word embeddings can be effectively harnessed for carving out attribute meaning from adjective-noun phrases. Observed improvements over the previous state-of-the-art are due to the type of word representation as such (dense neural embeddings vs. distributional count models) as well as a learned compositionality function based on a full additive model capitalizing on weight matrices for balancing the contributions of adjectives and nouns. Moreover, we were able to show that the compositionality function captures a generalized compositional process in the semantics of attribute-denoting adjective-noun phrases rather than the lexical meaning of individual attributes. Therefore, the proposed approach (i) poses an interesting alternative to previous distributional models which explicitly encode attribute meaning in word vectors and rely on vector mixture operations in order to compose them into attribute-based phrase representations, and (ii) bears the potential of being used as a generalized attribute extraction model on various domains of applications that demand for different attribute inventories.

5 Similarity Prediction Experiments

In this experiment, we assess the scalability of the previously trained composition models to different tasks by applying them to the prediction of semantic similarity in pairs of adjective-noun phrases.

5.1 Data

Our experiments are based on the adjective-noun section of the evaluation data set released by Mitchell and Lapata (2010). It consists of 108 pairs of adjective-noun phrases that were rated for similarity on a 7-point scale[7] by 54 human judges. In total, the data set comprises 1944 data points.

5.2 Experiment 3: Predicting Adjective-Noun Phrase Similarity

Experimental Procedure. For a given pair of adjective-noun phrases, we compute two phrase representations using word embeddings as word representations and compositionality functions trained on the HeiPLAS-Core subset, which achieved the best attribute selection results in Experiments 1 and 2. In the next step, we compute the cosine similarity between these two phrase representations. We correlate the results with human similarity ratings using Spearman's ρ and compare the resulting correlation scores to the reported results of Mitchell and Lapata (2010).

Baseline Models. We compare our models against the following approaches from the literature which were evaluated on the same data set: C-LDA (Hartung and Frank, 2011a), M&L-BoW and M&L-Topic (both by Mitchell and Lapata (2010)). All baseline models are count-based distributional models which differ in their underlying representation of word meaning: M&L-BoW relies on bag-of-words context windows, M&L-Topic and C-LDA use topics and attribute nouns as dimensions of meaning, respectively.

Results. As shown in Table 3, the best correlation scores between human similarity judgments and model predictions are achieved by our model that is built upon word embeddings and a trained full additive composition function based on weighting adjective and noun vectors (ρ=0.50). This model outperforms all distributional baseline models using vector mixtures as composition functions.

[7] A score of 1 expresses low similarity between phrases, 7 indicates high similarity.

Underlying Word Representation	$\odot$	$\oplus$	Weighted Addition	Full Additive
word2vec	0.36	0.48	0.42	**0.50**
M&L-BoW	0.46	0.36	0.44	n/a
M&L-Topic	0.25	0.37	0.38	n/a
C-LDA	0.28	0.19	n/a	n/a

Table 3: Results of Experiment 3 (Spearman's ρ between human judgments and model predictions)

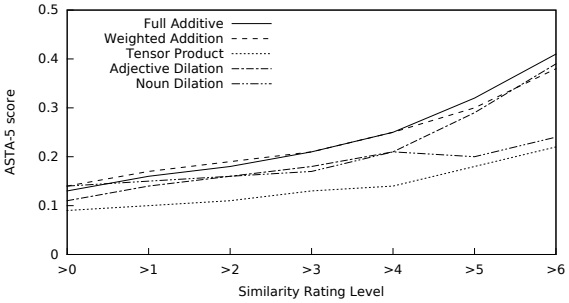

Figure 2: ASTA-5 scores over different levels of human similarity ratings (cf. Experiment 4)

With respect to weighted addition, all results reported in Table 3 are based on the weighting parameters (α=0.88; β=0.12) that have been found as optimal by Mitchell and Lapata (2010). Based on a grid search, we find α=0.60 and β=0.40 to be the best weighting parameters on our data. In this setting, the performance of the weighted vector addition model on word2vec embeddings can be increased to ρ=0.47, which is still slightly below unweighted vector addition on embeddings (ρ=0.48). Apparently, scalar weights in pointwise vector addition are quite sensitive to the underlying word representation. In the particular case of using word embeddings for similarity prediction, the contribution of the noun to the compositional semantics of the phrase seems to be relatively stronger than in the attribute selection task (cf. Experiment 1).

In total, these results indicate that compositionality functions optimized on the task of attribute selection can be effectively transferred to similarity prediction. This suggests that attribute meaning might be a prominent source of similarity in adjective-noun phrases, which will be subject to a closer investigation in the next experiment.

5.3 Experiment 4: Interpreting the Source of Similarity

Research in distributional semantics tends to focus on the *degree* of similarity between words or phrases, while the *source* of similarity is largely neglected (cf. Hartung (2015)). In this experiment, we hypothesize that attribute meaning provides a plausible explanation for the observed degree of similarity in phrase pairs from the M&L data set.

Experimental Procedure. For a given phrase pair, we compute the top-5 most similar attributes for each phrase in terms of their nearest neighbours in $\mathcal{S}$ (cf. Section 3.2). Then, both phrases are compared wrt. the proportion of shared attributes within these top-5 predictions. Averaging this score over all phrase pairs which were assigned a particular similarity rating by the human judges yields an *Average Shared Top-5 Attributes* (ASTA-5) score for this similarity level.

Results. Figure 2 plots ASTA-5 scores at different levels of human similarity ratings. We observe a general trend across all compositionality functions investigated: The higher the rating cutoff, the higher the number of shared attributes. Thus, with increasing similarity between two phrases (according to human ratings), the proportion of shared attributes in their compositional semantics tends to increase as well. Moreover, for highly similar pairs (rating cutoff$>$5), the full additive vector addition model yields the highest ASTA-5 scores.

Beyond this quantitative analysis, two of the authors manually investigated the shared attributes in 38 high-similarity phrase pairs (rating cutoff$>$4) as predicted by the weighted vector addition model wrt. their potential as plausible sources of similarity. We find that in 28 phrase pairs (73.6%), the predicted attribute is considered a plausible source of similarity, in eight others (26.4%), the predicted attribute does not explain the high similarity. The agreement between the annotators in terms of Fleiss' Kappa amounts to $\kappa = 0.62$.

5.4 Discussion

Our results show that a full additive compositional model trained to target attribute meaning improves performance on similarity prediction. This supports the interpretation that attributes are (at least)

a partial source of similarity between adjective-noun phrases. In fact, this has been corroborated by a preliminary manual investigation of shared attributes between high-similarity phrases. However, there is also evidence for several cases in which attribute meaning falls short of explaining high phrase similarity. This holds for phrases involving abstract concepts, for instance (cf. Hartung (2015), Borghi and Binkofski (2014)).

Nevertheless, we consider it a strength of our model that it is capable of providing plausible explanations in cases where attribute meaning is the most prominent source of similarity.

6 Conclusions

We have presented a model of attribute meaning in adjective-noun phrases that capitalizes on CBOW word embeddings. In our experiments, the model proves remarkably versatile as it advances the state-of-the-art in the two tasks of attribute selection and phrase similarity prediction. In the latter task, the property of being fully interpretable wrt. attributes as the potential source of similarities became apparent as an additional asset rendering the model potentially interoperable with knowledge representation formalisms and resources.

Improvements over previous distributional models can be traced back to two major sources: First, CBOW word embeddings work surprisingly well at the word level for capturing attribute meaning in adjectives (not for nouns, though). Future work should investigate whether further improvements can be obtained from more adjective-specific word embeddings that are trained on symmetric coordination patterns (Schwartz et al., 2016). Second, a learned compositionality function is effective at promoting attribute meaning in composed phrase representations. Best performances across both tasks are achieved by a full additive model with distinct weight matrices for the adjective and noun constituent. A trained tensor product that comes closer to the linguistic notion of functional application also performs well beyond the previous state-of-the-art, while falling short of the additive model. Apparently, more training data is needed to exhaust the full potential of the tensor product. Alternatively, tensor decomposition techniques along the lines of Shah et al. (2015) may be a possible way of coping with the large parameter space of the tensor approach.

Moreover, the learned compositionality function turns out to generalize well over individual attributes, which we consider a very promising result wrt. the suitability of the model in various NLP tasks such as aspect-based sentiment analysis. In future work, we are going to extend the present model to consider broader linguistic contexts and more varied syntactic configurations.

Acknowledgments

We gratefully acknowledge feedback and comments by the anonymous EACL reviewers, which considerably helped to improve the paper. This work was supported by the Cluster of Excellence *Cognitive Interaction Technology* 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG), and by the German Federal Ministry of Education and Research (BMBF) in the *KogniHome* project.

References

Abdulrahman Almuhareb. 2006. *Attributes in lexical acquisition*. Ph.D. thesis, University of Essex.

Omid Bakhshandeh and James F. Allen. 2015. From Adjective Glosses to Attribute Concepts: Learning Different Aspects That an Adjective Can Describe. In *Proceedings of the 11th International Conference on Computational Semantics (IWCS)*, pages 23–33, London, UK.

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

Marco Baroni, Raffaella Bernardi, and Roberto Zamparelli. 2014a. Frege in Space: A Program for Compositional Distributional Semantics. *Linguistic Issues in Language Technology*, 9:241–346.

Marco Baroni, Georgiana Dinu, and Germàn Kruszewski. 2014b. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In Kristina Toutanova and Hua Wu, editors, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland, June. Association for Computational Linguistics.

Lawrence W. Barsalou. 1992. Frames, Concepts and Conceptual Fields. In A. Lehrer and E.F. Kittay, editors, *Frames, Fields and Contrasts*, pages 21–74. Lawrence Erlbaum Associates, Hillsday, NJ.

Anna M. Borghi and Ferdinand Binkofski. 2014. *Words as Social Tools: An Embodied View on Abstract Concepts*. Springer Briefs in Cognition. Springer.

Antoine Bride, Tim Van de Cruys, and Nicholas Asher. 2015. A Generalisation of Lexical Functions for Composition in Distributional Semantics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 281–291, Beijing, China, July. Association for Computational Linguistics.

François Chollet. 2016. keras. `https://github.com/fchollet/keras`.

Philipp Cimiano. 2006. *Ontology Learning and Population from Text. Algorithms, Evaluation and Applications*. Springer.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Silvio Cordeiro, Carlos Ramisch, Marco Idiart, and Aline Villavicencio. 2016. Predicting the Compositionality of Nominal Compounds: Giving Word Embeddings a Hard Time. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1986–1997, Berlin, Germany, August. Association for Computational Linguistics.

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

Katrin Erk. 2016. What do you know about an alligator when you know the company it keeps? *Semantics & Pragmatics*, 9:1–63.

Gottlob Frege. 1892. Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100:25–50.

Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011. Experimental Support for a Categorical Compositional Distributional Model of Meaning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1394–1404. Association for Computiational Linguistics.

Edward Grefenstette, Mehrnoosh Sadrzadeh, Stephen Clark, Bob Coecke, and Stephen Pulman. 2014. Concrete Sentence Spaces for Compositional Distributional Models of Meaning. In Harry Bunt, Johan Bos, and Stephen Pulman, editors, *Computing Meaning*, volume 4, pages 71–86. Springer.

Nicola Guarino. 1992. Concepts, Attributes and Arbitrary Relations. *Data & Knowledge Engineering*, 8:249–261.

Emiliano Guevara. 2010. A regression model of adjective-noun compositionality in distributional semantics. In Roberto Basili and Marco Pennacchiotti, editors, *Proceedings of the 2010 Workshop on Geometrical Models of Natural Language Semantics*, pages 33–37, Uppsala, Sweden, July. Association for Computational Linguistics.

Zellig S. Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Matthias Hartung and Anette Frank. 2010. A Structured Vector Space Model for Hidden Attribute Meaning in Adjective-Noun Phrases. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING),* Beijing, China, pages 430–438.

Matthias Hartung and Anette Frank. 2011a. Assessing interpretable, attribute-related meaning representations for adjective-noun phrases in a similarity prediction task. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 52–61, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matthias Hartung and Anette Frank. 2011b. Exploring Supervised LDA Models for Assigning Attributes to Adjective-Noun Phrases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 540–551, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matthias Hartung. 2015. *Distributional Semantic Models of Attribute Meaning in Adjectives and Nouns*. Ph.D. thesis, Heidelberg University.

Diederik Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*.

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

Sebastian Löbner. 2013. *Understanding Semantics*. Routledge, 2nd edition.

Christopher D. Manning. 2015. Computational Linguistics and Deep Learning. *Computational Linguistics*, 41:701–707.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Efficient estimation of word representations in vector space. In *Proceedings of ICLR Workshop*.

George Miller and Christiane Fellbaum. 1998. Wordnet: An electronic lexical database.

Jeff Mitchell and Mirella Lapata. 2010. Composition in Distributional Models of Semantics. *Cognitive Science*, 34(8):1388–1429.

Mark Palatucci, Dean Pomerleau, Geoffrey Hinton, and Tom M. Mitchell. 2009. Zero-shot learning with semantic output codes. In *Proceedings of NIPS*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Wiebke Petersen and Oliver Hellwig. 2016. Exploring the value space of attributes: Unsupervised bidirectional clustering of adjectives in German. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2839–2848, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In René Witte, Hamish Cunningham, Jon Patrick, Elena Beisswanger, Ekaterina Buyko, Udo Hahn, Karin Verspoor, and Anni R. Coden, editors, *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA.

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

Roy Schwartz, Roi Reichart, and Ari Rappoport. 2016. Symmetric patterns and coordinations: Fast and enhanced representations of verbs and adjectives. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 499–505, San Diego, California, June. Association for Computational Linguistics.

Parikshit Shah, Nikhil Rao, and Gongguo Tang. 2015. Sparse and low-rank tensor decomposition. In *Proceedings of NIPS*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October. Association for Computational Linguistics.

Niket Tandon, Gerard de Melo, Fabian Suchanek, and Gerhard Weikum. 2014. WebChild: Harvesting and Organizing Commonsense Knowledge from the Web. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, pages 523–532, New York, NY, USA. ACM.

Dominic Widdows. 2008. Semantic Vector Products: Some Initial Investigations. In *Proceedings of the 2nd Conference on Quantum Interaction*, Oxford, UK.