A Structured Vector Space Model for Hidden Attribute Meaning in Adjective-Noun Phrases

Matthias Hartung and **Anette Frank**
Computational Linguistics Department
Heidelberg University
{hartung, frank}@cl.uni-heidelberg.de

Abstract

We present an approach to model hidden attributes in the compositional semantics of adjective-noun phrases in a distributional model. For the representation of *adjective meanings*, we reformulate the pattern-based approach for attribute learning of Almuhareb (2006) in a structured vector space model (VSM). This model is complemented by a structured vector space representing attribute dimensions of *noun meanings*. The combination of these representations along the lines of compositional semantic principles exposes the underlying semantic relations in adjective-noun phrases. We show that our compositional VSM outperforms simple pattern-based approaches by circumventing their inherent sparsity problems.

1 Introduction

In formal semantic theory, the compositional semantics of adjective-noun phrases can be modeled in terms of *selective binding* (Pustejovsky, 1995), i.e. the adjective selects one of possibly several roles or attributes[1] from the semantics of the noun.

(1) a. a blue car
b. COLOR(car)=blue

In this paper, we define a distributional framework that models the compositional process underlying the modification of nouns by adjectives. We focus on property-denoting adjectives as they are valuable for acquiring concept representations for, e.g., ontology learning. An approach for automatic subclassification of property-denoting adjectives is presented in Hartung and Frank (2010). Our goal is to expose, for adjective-noun phrases as in (1a), the attribute in the semantics of the noun that is selected by the adjective, while not being overtly realized on the syntactic level. The semantic information we intend to capture for (1a) is formalized in (1b).

Ideally, this kind of knowledge could be extracted from corpora by searching for patterns that paraphrase (1a), e.g. *the color of the car is blue*. However, linguistic patterns that explicitly relate nouns, adjectives and attributes are very rare.

We avoid these sparsity issues by reducing the triple $r{=}\langle noun, attribute, adjective\rangle$ that encodes the relation illustrated in (1b) to tuples $r'{=}\langle noun, attribute\rangle$ and $r''{=}\langle attribute, adjective\rangle$, as suggested by Turney and Pantel (2010) for similar tasks. Both r' and r'' can be observed much more frequently in text corpora than r. Moreover, this enables us to model adjective and noun meanings as distinct semantic vectors that are built over attributes as dimensions. Based on these semantic representations, we make use of vector composition operations in order to reconstruct r from r' and r''. This, in turn, allows us to infer complete noun-attribute-adjective *triples* from individually acquired noun-attribute and adjective-attribute representations.

The contributions of our work are as follows: (i) We propose a framework for attribute selection based on structured vector space models (VSM), using as meaning dimensions attributes elicited

[1] In the original statement of the theory, adjectives select *qualia roles* that can be considered as collections of attributes.

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 430–438,
Beijing, August 2010

by adjectives; (ii) we complement this novel representation of adjective meaning with structured vectors for *noun meanings* similarly built on attributes as meaning dimensions; (iii) we propose a composition of these representations that mirrors principles of compositional semantics in mapping adjective-noun phrases to their corresponding ontological representation; (iv) we propose and evaluate several metrics for the selection of meaningful components from vector representations.

2 Related Work

Adjective-noun meaning composition has not been addressed in a distributional framework before (cf. Mitchell and Lapata (2008)). Our approach leans on related work on attribute learning for ontology induction and recent work in distributional semantics.

Attribute learning. Early approaches to attribute learning include Hatzivassiloglou and McKeown (1993), who cluster adjectives that denote values of the same attribute. A weakness of their work is that the type of the attribute cannot be made explicit. More recent attempts to attribute learning from adjectives are Cimiano (2006) and Almuhareb (2006). Cimiano uses attributes as features to arrange sets of concepts in a lattice. His approach to attribute acquisition harnesses adjectives that occur frequently as concept modifiers in corpora. The association of adjectives with their potential attributes is performed by dictionary look-up in WordNet (Fellbaum, 1998). Similarly, Almuhareb (2006) uses adjectives and attributes as (independent) features for the purpose of concept learning. He acquires adjective-attribute pairs using a pattern-based approach.

As a major limitation, these approaches are confined to adjective-attribute pairs. The polysemy of adjectives that can only be resolved in the context of the modified noun is entirely neglected.

From a methodological point of view, our work is similar to Almuhareb's, as we will also build on lexico-syntactic patterns for attribute selection. However, we extend the task to involve nouns and rephrase his approach in a distributional framework based on the composition of structured vector representations.

Distributional semantics. We observe two recent trends in distributional semantics research: (i) The use of VSM tends to shift from measuring unfocused semantic similarity to capturing increasingly fine-grained semantic information by incorporating more linguistic structure. Following Baroni and Lenci (to appear), we refer to such models as *structured vector spaces*. (ii) Distributional methods are no longer confined to word meaning, but are noticeably extended to capture meaning on the *phrase level*. Prominent examples for (i) are Padó and Lapata (2007) and Rothenhäusler and Schütze (2009) who use syntactic dependencies rather than single word co-occurrences as dimensions of semantic spaces. Erk and Padó (2008) extend this idea to the argument structure of verbs, while also accounting for compositional meaning aspects by modelling predication over arguments. Hence, their work is also representative for (ii).

Baroni et al. (2010) use lexico-syntactic patterns to represent concepts in a structured VSM whose dimensions are interpretable as empirical manifestations of properties. We rely on similar techniques for the acquisition of structured vectors, whereas our work focusses on exposing the hidden meaning dimensions involved in compositional processes underlying concept modification.

The commonly adopted method for modelling compositionality in VSM is vector composition (Mitchell and Lapata, 2008; Widdows, 2008). Showing the benefits of vector composition for language modelling, Mitchell and Lapata (2009) emphasize its potential to become a standard method in NLP.

The approach pursued in this paper builds on both lines of research sketched in (i) and (ii) in that we model a specific meaning layer in the semantics of adjectives and nouns in a structured VSM. Vector composition is used to expose their hidden meaning dimensions on the phrase level.

3 Structured Vector Representations for Adjective-Noun Meaning

3.1 Motivation

Contrary to prior work, we model attribute selection as involving *triples* of nouns, attributes and

	COLOR	DIRECTION	DURATION	SHAPE	SIZE	SMELL	SPEED	TASTE	TEMPERATURE	WEIGHT
v_e	1	1	0	1	45	0	4	0	0	21
v_b	14	38	2	20	26	0	45	0	0	20
$v_e \times v_b$	14	38	0	20	**1170**	0	180	0	0	420
$v_e + v_b$	15	39	2	21	**71**	0	49	0	0	41

Figure 1: Vectors for *enormous* (v_e) and *ball* (v_b)

adjectives, as in (2). The triple r can be broken down into tuples $r' = \langle noun, attribute \rangle$ and $r'' = \langle attribute, adjective \rangle$. Previous learning approaches focussed on r' (Cimiano, 2006) or r'' (Almuhareb, 2006) only.

(2) a. a blue$_{\text{value}}$ car$_{\text{concept}}$
b. ATTR(concept) = value

In semantic composition of adjective-noun compounds, the adjective (e.g. *blue*) contributes a value for an attribute (here: COLOR) that characterizes the concept evoked by the noun (e.g. *car*). Thus, the attribute in (2) constitutes a 'hidden variable' that is not overtly expressed in (2a), but constitutes the central axis that relates r' and r''.

Structured vectors built on extraction patterns. We model the semantics of adjectives and nouns in a structured VSM that conveys the hidden relationship in (2). The dimensions of the model are defined by attributes, such as COLOR, SIZE or SPEED, while the vector components are determined on the basis of carefully selected acquisition patterns that are tailored to capturing the particular semantic information of interest for r' and r''. In this respect, lexico-syntactic patterns serve a similar purpose as dependency relations in Padó and Lapata (2007) or Rothenhäusler and Schütze (2009). The upper part of Fig. 1 displays examples of vectors we build for adjectives and nouns.

Composing vectors along hidden dimensions. The fine granularity of lexico-syntactic patterns that capture the triple r comes at the cost of their sparsity when applied to corpus data. Therefore, we construct separate vector representations for r' and r''. Eventually, these representations are joined by vector composition to reconstruct the triple r. Apart from avoiding sparsity issues, this compositional approach has several prospects from a linguistic perspective as well.

Ambiguity and disambiguation. Building vectors with attributes as meaning dimensions enables us to model (i) ambiguity of adjectives with regard to the attributes they select, and (ii) the disambiguation capacity of adjective and noun vectors when considered jointly. Consider, for example, the phrase *enormous ball* that is ambiguous for two reasons: *enormous* may select a set of possible attributes (SIZE or WEIGHT, among others), while *ball* elicits several attributes in accordance with its different word senses[2]. As seen in Fig. 1, these ambiguities are nicely captured by the separate vector representations for the adjective and the noun (upper part); by composing these representations, the ambiguity is resolved (lower part).

3.2 Building a VSM for Adjective-Noun Meaning

In this section, we introduce the methods we apply in order to (i) acquire vector representations for adjectives and nouns, (ii) select appropriate attributes from them, and (iii) compose them.

3.2.1 Attribute Acquisition Patterns

We use the following patterns[3] for the acquisition of vectors capturing the tuple $r'' = \langle attribute, adjective \rangle$. Even though some of these patterns (A1 and A4) match triples of nouns, attributes and adjectives, we only use them for the extraction of binary tuples (underlined), thus abstracting from the modified noun.

(A1) ATTR of DT? NN is|was JJ
(A2) DT? RB? JJ ATTR
(A3) DT? JJ or JJ ATTR
(A4) DT? NN's ATTR is|was JJ
(A5) is|was|are|were JJ in|of ATTR

To acquire noun vectors capturing the tuple $r' = \langle noun, attribute \rangle$, we rely on the following patterns. Again, we only extract pairs, as indicated by the underlined elements.

(N1) NN with|without DT? RB? JJ? ATTR
(N2) DT ATTR of DT? RB? JJ? NN
(N3) DT NN's RB? JJ? ATTR
(N4) NN has|had a|an RB? JJ? ATTR

[2] WordNet senses for the noun *ball* include, among others: 1. *round object [...] in games*; 2. *solid projectile*, 3. *object with a spherical shape*, 4. *people [at a] dance*.

[3] Some of these patterns are taken from Almuhareb (2006) and Sowa (2000). The descriptions rely on the Penn Tagset (Marcus et al., 1999). ? marks optional elements.

3.2.2 Target Filtering

Some of the adjectives extracted by A1-A5 are not property-denoting and thus represent noise. This affects in particular pattern A2, which extracts adjectives like *former* or *more*, or relational ones such as *economic* or *geographic*.

This problem may be addressed in different ways: By *target filtering*, extractions can be checked against a predicative pattern P1 that is supposed to apply to property-denoting adjectives only. Vectors that fail this test are suppressed.

(P1) DT NN is|was JJ

Alternatively, extractions obtained from low-confidence patterns can be awarded reduced weights by means of a *pattern value function* (defined in 3.3; cf. Pantel and Pennacchiotti (2006)).

3.2.3 Attribute Selection

We intend to use the acquired vectors in order to detect attributes that are implicit in adjective-noun meaning. Therefore, we need a method that selects appropriate attributes from each vector. While, in general, this task consists in distinguishing semantically meaningful dimensions from noise, the requirements are different depending on whether attributes are to be selected from adjective or noun vectors. This is illustrated in Fig. 1, a typical configuration, with one vector representing a typical property-denoting adjective that exhibits relatively strong peaks on one or more dimensions, whereas noun vectors show a tendency for broad and flat distributions over their dimensions. This suggests using a strict selection function (choosing few very prominent dimensions) for adjectives and a less restrictive one (licensing the inclusion of more dimensions of lower relative prominence) for nouns. Moreover, we are interested in finding a selection function that relies on as few free parameters as possible in order to avoid frequency or dimensionality effects.

MPC Selection (MPC). An obvious method for attribute selection is to choose the most prominent component from any vector (i.e., the highest absolute value). If a vector exhibits several peaks, all other components are rejected, their relative importance notwithstanding. MPC obviously fails to capture polysemy of targets, which affects adjectives such as *hot*, in particular.

Threshold Selection (TSel). TSel recasts the approach of Almuhareb (2006), in selecting all dimensions as attributes whose components exceed a frequency threshold. This avoids the drawback of MPC, but introduces a parameter that needs to be optimized. Also, it is difficult to apply absolute thresholds to composed vectors, as the range of their components is subject to great variation, and it is unclear whether the method will scale with increased dimensionality.

Entropy Selection (ESel). In information theory, entropy measures the average uncertainty in a probability distribution (Manning and Schütze, 1999). We define the entropy $H(v)$ of a vector $v{=}\langle v_1, \ldots, v_n \rangle$ over its components as $H(v) = -\sum_{i=1}^{n} P(v_i) \log P(v_i)$, where $P(v_i) = v_i / \sum_{i=1}^{n} v_i$.

We use $H(v)$ to assess the impact of singular vector components on the overall entropy of the vector: We expect entropy to detect components that contribute noise, as opposed to those that contribute important information.

We define an algorithm for entropy-based attribute selection that returns a list of informative dimensions. The algorithm successively suppresses (combinations of) vector components one by one. Given that a gain of entropy is equivalent to a loss of information and vice versa, we assume that every combination of components that leads to an increase in entropy when being suppressed is actually responsible for a substantial amount of information. The algorithm includes a back-off to MPC for the special case that a vector contains a single peak (i.e., $H(v) = 0$), so that, in principle, it should be applicable to vectors of any kind. Vectors with very broad distributions over their dimensions, however, pose a problem to this method. For *ball* in Fig. 1, for instance, the method does not select any dimension.

Median Selection (MSel). As a further method we rely on the median m that can be informally defined as the value that separates the upper from the lower half of a distribution (Krengel, 2003). It is less restrictive than MPC and TSel and overcomes the particular drawback of ESel. Using this measure, we choose all dimensions whose components exceed m. Thus, for the vector representing

Pattern Label	# Hits (Web)	# Hits (ukWaC)
A1	2249	815
A2	36282	72737
A3	3370	1436
A4	–	7672
A5	–	3768
N1	–	682
N2	–	5073
N3	–	953
N4	–	56

Table 1: Number of pattern hits on the Web (Almuhareb, 2006) and on ukWaC

ball, WEIGHT, DIRECTION, SHAPE, SPEED and SIZE are selected.

3.2.4 Vector Composition

We use vector composition as a hinge to combine adjective and noun vectors in order to reconstruct the triple $r{=}\langle noun, attribute, adjective\rangle$. Mitchell and Lapata (2008) distinguish two major classes of vector composition operations, namely multiplicative and additive operations, that can be extended in various ways. We use their standard definitions (denoted $\times$ and $+$, henceforth). For our task, we expect $\times$ to perform best as it comes closest to the linguistic function of *intersective* adjectives, i.e. to select dimensions that are prominent both for the adjective and the noun, whereas $+$ basically blurs the vector components, as can be seen in the lower part of Fig. 1.

3.3 Model Parameters

We follow Padó and Lapata (2007) in defining a semantic space as a matrix $M = B \times T$ relating a set of target elements T to a set of basis elements B. Further parameters and their instantiations we use in our model are described below. We use p to denote an individual lexico-syntactic pattern.

The **basis elements** of our VSM are nouns denoting attributes. For comparison, we use the attributes selected by Almuhareb (2006): COLOR, DIRECTION, DURATION, SHAPE, SIZE, SMELL, SPEED, TASTE, TEMPERATURE, WEIGHT.

The **context selection function** *cont(t)* determines the set of patterns that contribute to the representation of each target word $t \in T$. These are the patterns A1-A5 and N1-N4 (cf. Section 3.2.1).

The **target elements** represented in the vector space comprise all adjectives T_A that match the patterns A1 to A5 in the corpus, provided they exceed a frequency threshold n. During development, n was set to 5 in order to filter noise.

As for the target nouns T_N, we rely on a representative dataset compiled by Almuhareb (2006). It contains 402 nouns that are balanced with regard to semantic class (according to the WordNet supersenses), ambiguity and frequency.

As **association measure** that captures the strength of the association between the elements of B and T, we use raw frequency counts[4] as obtained from the PoS-tagged and lemmatized version of the ukWaC corpus (Baroni et al., 2009). Table 1 gives an overview of the number of hits returned by these patterns.

The **basis mapping function** μ creates the dimensions of the semantic space by mapping each extraction of a pattern p to the attribute it contains.

The **pattern value function** enables us to subdivide dimensions along particular patterns. We experimented with two instantiations: pv_{const} considers, for each dimension, all patterns, while weighting them equally. $pv_f(p)$ awards the extractions of pattern p with weight 1, while setting the weights for all patterns different from p to 0.

4 Experiments

We evaluate the performance of the structured VSM on the task of inferring attributes from adjective-noun phrases in three experiments: In Exp1 and Exp2, we evaluate vector representations capturing r' and r'' independently of one another. Exp3 investigates the selection of hidden attributes from vector representations constructed by composition of adjective and noun vectors.

We compare all results against different *gold standards*. In Exp1, we follow Almuhareb (2006), evaluating against WordNet 3.0. For Exp2 and Exp3, we establish gold standards manually: For Exp2, we construct a test set of nouns annotated with their corresponding attributes. For Exp3, we manually annotate adjective-noun phrases with the attributes appropriate for the whole phrase. All experiments are evaluated in terms of precision, recall and F_1 score.

[4] We experimented with the conditional probability ratio proposed by Mitchell and Lapata (2009). As it performed worse on our data, we did not consider it any further.

4.1 Exp1: Attribute Selection for Adjectives

The first experiment evaluates the performance of structured vector representations on attribute selection for adjectives. We compare this model against a re-implementation of Almuhareb (2006).

Experimental settings and gold standard. To reconstruct Almuhareb's approach, we ran his patterns A1-A3 on the ukWaC corpus. Table 1 shows the number of hits when applied to the Web (Almuhareb, 2006) vs. ukWaC. A1 and A3 yield less extractions on ukWaC as compared to the Web.[5] We introduced two additional patterns, A4 and A5, that contribute about 10,000 additional hits. We adopted Almuhareb's manually chosen thresholds for attribute selection for A1-A3; for A4, A5 and a combination of all patterns, we manually selected optimal thresholds.

We experiment with pv_{const} and all variants of $pv_f(p)$ for pattern weighting (see sect. 3.3). For attribute selection, we compare TSel (as used by Almuhareb), ESel and MSel.

The gold standard consists of all adjectives that are linked to at least one of the ten attributes we consider by WordNet's `attribute` relation (1063 adjectives in total).

Evaluation results. Results for Exp1 are displayed in Table 2. The settings of pv are given in the rows, the attribute selection methods (in combination with target filtering[6]) in the columns.

The results for our re-implementation of Almuhareb's individual patterns are comparable to his original figures[7], except for A3 that seems to suffer from quantitative differences of the underlying data. Combining all patterns leads to an improvement in precision over (our reconstruction of) Almuhareb's best individual pattern when TSel and target filtering are used in combination. MPC and MSel perform worse (not reported here). As for target filtering, A1 and A3 work best.

Both TSel and ESel benefit from the combination with the target filter, where the largest improvement (and the best overall result) is observable for ESel on pattern A1 only. This is the pattern that performs worst in Almuhareb's original setting. From this, we conclude that both ESel and target filtering are valuable extensions to pattern-based structured vector spaces if precision is in focus. This also underlines a finding of Rothenhäusler and Schütze (2009) that VSMs intended to convey specific semantic information rather than mere similarity benefit primarily from a linguistically adequate choice of contexts.

Similar to Almuhareb, recall is problematic. Even though ESel leads to slight improvements, the scores are far from satisfying. With Almuhareb, we note that this is mainly due to a high number of extremely fine-grained adjectives in WordNet that are rare in corpora.[8]

	MPC			ESel			MSel		
	P	R	F	P	R	F	P	R	F
pv_f(N1)	0.22	0.06	0.10	**0.29**	0.04	0.07	0.22	**0.09**	**0.13**
pv_f(N2)	**0.29**	0.18	0.23	0.20	0.06	0.09	0.28	**0.39**	**0.33**
pv_f(N3)	**0.34**	0.05	0.09	0.20	0.02	0.04	0.25	**0.08**	**0.12**
pv_f(N4)	0.25	0.02	0.04	**0.29**	0.02	0.03	0.26	0.02	**0.05**
pv_{const}	**0.29**	0.18	0.22	0.20	0.06	0.09	0.28	**0.43**	**0.34**

Table 3: Evaluation results for Experiment 2

4.2 Exp2: Attribute Selection for Nouns

Exp2 evaluates the performance of attribute selection from noun vectors tailored to the tuple r''.

Construction of the gold standard. For evaluation, we created a gold standard by manually annotating a set of nouns with attributes. This gold standard builds on a random sample extracted from T_N (cf. section 3.3). Running N1-N4 on ukWaC returned semantic vectors for 216 concepts. From these, we randomly sampled 100 concepts that were manually annotated by three human annotators.

The annotators were provided a matrix consisting of the nouns and the set of ten attributes for each noun. Their task was to remove all inappropriate attributes. They were free to decide how many attributes to accept for each noun. In order to deal with word sense ambiguity, the annotators were instructed to consider all senses of a noun and to retain every attribute that was acceptable for at least one sense.

Inter-annotator agreement amounts to κ= 0.69 (Fleiss, 1971). Cases of disagreement were adjudicated by majority-voting. The gold standard

[5] The difference for A2 is an artifact of Almuhareb's extraction methodology.

[6] Regarding target filtering, we only report the best filter pattern for each configuration.

[7] P(A1)=0.176, P(A2)=0.218, P(A3)=0.504

	Almuhareb (reconstr.)				VSM (TSel + Target Filter)					VSM (ESel)			VSM (ESel + Target Filter)			
	P	R	F	Thr	P	R	F	Patt	Thr	P	R	F	P	R	F	Patt
$pv_f(\texttt{A1}) = 1$	0.183	0.005	0.009	5	0.300	0.004	0.007	A3	5	0.231	**0.045**	**0.076**	**0.519**	0.035	0.065	A3
$pv_f(\texttt{A2}) = 1$	0.207	0.039	0.067	50	**0.300**	0.033	0.059	A1	50	0.084	**0.136**	**0.104**	0.240	0.049	0.081	A3
$pv_f(\texttt{A3}) = 1$	0.382	0.020	0.039	5	**0.403**	0.014	0.028	A1	5	0.192	**0.059**	**0.090**	0.375	0.027	0.050	A1
$pv_f(\texttt{A4}) = 1$					**0.301**	0.020	0.036	A3	10	0.135	**0.055**	**0.078**	0.272	0.020	0.038	A1
$pv_f(\texttt{A5}) = 1$					0.295	0.008	0.016	A3	24	0.105	**0.056**	**0.073**	**0.315**	0.024	0.045	A3
pv_{const}					**0.420**	0.024	0.046	A1	183	0.076	**0.152**	**0.102**	0.225	0.054	0.087	A3

Table 2: Evaluation results for Experiment 1

contains 424 attributes for 100 nouns.

Evaluation results. Results for Exp2 are given in Table 3. Performance is lower in comparison to Exp1. We hypothesize that the tuple r'' might not be fully captured by overt linguistic patterns. This needs further investigation in future research.

Against this background, MPC is relatively precise, but poor in terms of recall. ESel, being designed to select more than one prominent dimension, counterintuitively fails to increase recall, suffering from the fact that many noun vectors show a rather flat distribution without any strong peak. MSel turns out to be most suitable for this task: Its precision is comparable to MPC (with `N3` as an outlier), while recall is considerably higher. Overall, these results indicate that attribute selection for adjectives and nouns, though similar, should be viewed as distinct tasks that require different attribute selection methods.

4.3 Exp3: Attribute Selection for Adjective-Noun Phrases

In this experiment, we compose noun and adjective vectors in order to yield a new combined representation. We investigate whether the semantic information encoded by the components of this new vector is sufficiently precise to disambiguate the attribute dimensions of the original representations (see section 3.1) and, thus, to infer hidden attributes from adjective-noun phrases (see (2)) as advocated by Pustejovsky (1995).

Construction of the gold standard. For evaluation, we created a manually annotated test set of adjective-noun phrases. We selected a subset of property-denoting adjectives that are appropriate modifiers for the nouns from T_N using the predicative pattern `P1` (see sect. 3) on ukWaC. This yielded 2085 adjective types that were further reduced to 386 by frequency filtering ($n = 5$). We sampled our test set from all pairs in the cartesian product of the 386 adjectives and 216 nouns (cf. Exp2) that occurred at least 5 times in a subsection of ukWaC. To ensure a sufficient number of ambiguous adjectives in the test set, sampling proceeded in two steps: First, we sampled four nouns each for a manual selection of 15 adjectives of all ambiguity levels in WordNet. This leads to 60 adjective-noun pairs. Second, another 40 pairs were sampled fully automatically.

The test set was manually annotated by the same annotators as in Exp2. They were asked to remove all attributes that were not appropriate for a given adjective-noun pair, either because it is not appropriate for the noun or because it is not selected by the adjective. Further instructions were as in Exp2, in particular regarding ambiguity.

The overall agreement is κ=0.67. After adjudication by majority voting, the resulting gold standard contains 86 attributes for 76 pairs. 24 pairs could not be assigned any attribute, either because the adjective did not denote a property, as in *private investment*, or the most appropriate attribute was not offered, as in *blue day* or *new house*.

We evaluate the vector composition methods discussed in section 3.2.4. Individual vectors for the adjectives and nouns from the test pairs were constructed using all patterns `A1`-`A5` and `N1`-`N4`. For attribute selection, we tested MPC, ESel and MSel. The results are compared against three baselines: BL-P implements a purely pattern-based method, i.e. running the patterns that extract the triple r (`A1`, `A4`, `N1`, `N3` and `N4`, with `JJ` and `NN` instantiated accordingly) on the pairs from the test set. BL-N and BL-Adj are back-offs for vector composition, taking the respective noun or adjective vector, as investigated in Exp1 and Exp2, as surrogates for a composed vector.

[8] For instance: *bluish-lilac*, *chartreuse* or *pink-lavender* as values of the attribute COLOR.

	MPC			ESel			MSel		
	P	R	F	P	R	F	P	R	F
$\times$	0.60	0.58	**0.59**	**0.63**	0.46	0.54	0.27	0.72	0.39
$+$	0.43	0.55	0.48	0.42	0.51	0.46	0.18	**0.91**	0.30
BL-Adj	0.44	0.60	0.50	0.51	0.63	0.57	0.23	0.83	0.36
BL-N	0.27	0.35	0.31	0.37	0.29	0.32	0.17	0.73	0.27
BL-P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4: Evaluation results for Experiment 3

Evaluation results. Results are given in Table 4. Attribute selection based on the composition of adjective and noun vectors yields a considerable improvement of both precision and recall as compared to the individual results obtained in Exp1 and Exp2. Comparing the results of Exp3 against the baselines reveals two important aspects of our work. First, the complete failure of BL-P[9] underlines the attractiveness of our method to build structured vector representations from patterns of reduced complexity. Second, vector composition is suitable for selecting hidden attributes from adjective-noun phrases that are jointly encoded by adjective and noun vectors: Both composition methods we tested outperform BL-N.

However, the choice of the composition method matters: $\times$ performs best with a maximum precision of 0.63. This confirms our expectation that vector multiplication is a good approximation for attribute selection in adjective-noun semantics. Being outperformed by BL-Adj in most categories, $+$ is less suited for this task.

All selection methods outperform BL-Adj in precision. Comparing MPC and ESel, ESel achieves better precision when combined with the $\times$-operator, while doing worse for recall. The robust performance of MPC is not surprising as the test set contains only ten adjective-noun pairs that are still ambiguous with regard to the attributes they elicit. The stronger performance of the entropy-based method with the $\times$-operator is mainly due to its accuracy on detecting false positives, in that it is able to return "empty" selections. In terms of precision, MSel did worse in general, while recall is decent. This underlines that vector composition generally promotes meaningful components, but MSel is too inaccurate to select them.

Given the performance of the baselines and the noun vectors in Exp2, we consider this a very promising result for our approach to attribute selection from structured vector representations. The results also corroborate the insufficiency of previous approaches to attribute learning from adjectives alone.

[9]The patterns used yield no hits for the test pairs at all.

5 Conclusions and Outlook

We proposed a structured VSM as a framework for inferring hidden attributes from the compositional semantics of adjective-noun phrases.

By reconstructing Almuhareb (2006), we showed that structured vector representations of adjective meaning consistently outperform simple pattern-based learning, up to 13 pp. in precision. A combination of target filtering and pattern weighting turned out to be effective here, by selecting particulary meaningful lexico-syntactic contexts and filtering adjectives that are not property-denoting. Further studies need to investigate this phenomenon and its most appropriate formulation in a vector space framework.

Moreover, the VSM offers a natural representation for sense ambiguity of adjectives. Comparing attribute selection methods on adjective and noun vectors shows that they are sensitive to the distributional structure of the vectors, and need to be chosen with care. Future work will investigate these selection methods in high-dimensional vectors spaces, by using larger sets of attributes.

Exp3 shows that the composition of pattern-based adjective and noun vectors robustly reflects aspects of meaning composition in adjective-noun phrases, with attributes as a hidden dimension. It also suggests that composition is effective in disambiguation of adjective and noun meanings. This hypothesis needs to be substantiated in further experiments.

Finally, we showed that composition of vectors representing complementary meaning aspects can be beneficial to overcome sparsity effects. However, our compositional approach meets its limits if the patterns capturing adjective and noun meaning in isolation are too sparse to acquire sufficiently populated vector components from corpora. For future work, we envisage using vector similarity to acquire structured vectors for infrequent targets from semantic spaces that convey less linguistic structure to address these remaining sparsity issues.

References

Almuhareb, Abdulrahman. 2006. *Attributes in Lexical Acquisition*. Ph.D. Dissertation, Department of Computer Science, University of Essex.

Baroni, Marco and Alessandro Lenci. to appear. Distributional Memory. A General Framework for Corpus-based Semantics. *Computational Linguistics*.

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

Baroni, Marco, Brian Murphy, Eduard Barbu, and Massimo Poesio. 2010. Strudel. A Corpus-based Semantic Model of Based on Properties and Types. *Cognitive Science*, 34:222–254.

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

Erk, Katrin and Sebastian Padó. 2008. A Structured Vector Space Model for Word Meaning in Context. In *Proceedings of EMNLP*, Honolulu, HI.

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

Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Hartung, Matthias and Anette Frank. 2010. A Semi-supervised Type-based Classification of Adjectives. Distinguishing Properties and Relations. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, Valletta, Malta, May.

Hatzivassiloglou, Vasileios and Kathleen McKeown. 1993. Towards the Automatic Identification of Adjectival Scales. Clustering Adjectives According to Meaning. In *Proceedings of the 31st Annual Meeting of the Association of Computational Linguistics*, pages 172–182.

Krengel, Ulrich. 2003. *Wahrscheinlichkeitstheorie und Statistik*. Vieweg, Wiesbaden.

Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

Marcus, Mitchell P., Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1999. Treebank-3, ldc99t42. CD-ROM. Philadelphia, Penn.: Linguistic Data Consortium.

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

Mitchell, Jeff and Mirella Lapata. 2009. Language Models Based on Semantic Composition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, August 2009, pages 430–439, Singapore, August.

Padó, Sebastian and Mirella Lapata. 2007. Dependency-based Construction of Semantic Space Models. *Computational Linguistics*, 33:161–199.

Pantel, Patrick and Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, 17–21 July 2006, pages 113–120.

Pustejovsky, James. 1995. *The Generative Lexicon*. MIT Press, Cambridge, Mass.

Rothenhäusler, Klaus and Hinrich Schütze. 2009. Unsupervised Classification with Dependency Based Word Spaces. In *Proceedings of the EACL Workshop on Geometrical Models of Natural Language Semantics (GEMS)*, pages 17–24, Athens, Greece, March.

Sowa, John F. 2000. *Knowledge Representation. Logical, Philosophical, and Computational Foundations*. Brooks Cole.

Turney, Peter D. and Patrick Pantel. 2010. From Frequency to Meaning. Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

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