Exploring Supervised LDA Models for Assigning Attributes to Adjective-Noun Phrases

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

Abstract

This paper introduces an attribute selection task as a way to characterize the inherent meaning of property-denoting adjectives in adjective-noun phrases, such as e.g. *hot* in *hot summer* denoting the attribute TEMPERATURE, rather than TASTE. We formulate this task in a vector space model that represents adjectives and nouns as vectors in a semantic space defined over possible attributes. The vectors incorporate latent semantic information obtained from two variants of LDA topic models. Our LDA models outperform previous approaches on a small set of 10 attributes with considerable gains on sparse representations, which highlights the strong smoothing power of LDA models. For the first time, we extend the attribute selection task to a new data set with more than 200 classes. We observe that large-scale attribute selection is a hard problem, but a subset of attributes performs robustly on the large scale as well. Again, the LDA models outperform the VSM baseline.

1 Introduction

Corpus-based statistical modeling of semantics is gaining increased attention in computational linguistics. This field of research includes distributional vector space models (VSMs), i.e., models that represent the semantics of words or phrases as vectors over high-dimensional cooccurrence data (Turney and Pantel, 2010; Baroni and Lenci, 2010, i.a.), as well as latent variable models (LVMs) which aggregate distributional observations in 'hidden', or latent variables, thereby reducing the dimensionality of the data. An example of the latter are topic models (Blei et al., 2003), which have recently been applied to modeling selectional preferences of verbs (Ritter et al., 2010; Ó Séaghdha, 2010), or word sense disambiguation (Li et al., 2010).

A topic that is increasingly studied in distributional semantics is the semantics of adjectives, both in isolation (Almuhareb, 2006) and in compositional adjective-noun phrases (Hartung and Frank, 2010; Guevara, 2010; Baroni and Zamparelli, 2010).

In this paper, we propose a new approach to a problem we denote as *attribute selection*: The task is to predict the hidden attribute meaning expressed by a property-denoting adjective in composition with a noun. The adjective *hot*, e.g., may denote attributes such as TEMPERATURE, TASTE or EMOTIONALITY. These adjective meanings can be combined with nouns such as *tea, soup* or *debate*, which can be characterized in terms of attributes as well. The goal of the task is to determine the hidden attribute meaning predicated over the noun in a given adjective-noun phrase, as illustrated in (1).

(1) a. a $\text{hot}_{\text{value}}$ $\text{summer}_{\text{concept}}$
b. TEMPERATURE(*summer*) = *hot*

It is by way of the composition of adjective and noun that specific attributes are selected from the adjective's space of possible attribute meanings, and typically lead to a disambiguation of the adjective and possibly the noun. Hartung and Frank (2010) were the first to model this insight in a VSM by representing the meaning of adjectives and nouns in semantic vectors defined over attributes. The meaning of adjective-noun phrases is computed by means of

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 540–551,
Edinburgh, Scotland, UK, July 27–31, 2011.

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

Figure 1: Vectors for *enormous* ($\vec{e}$) and *ball* ($\vec{b}$)

vector composition, such that the 'hidden' attribute meaning of the phrase can be 'selected' as a prominent component from the composed vector. This is illustrated in Fig. 1 for the adjective *enormous* ($\vec{e}$) in combination with the noun *ball* ($\vec{b}$), with alternative composition operations: vector multiplication ($\times$) and addition ($+$).[1] Both yield SIZE as the most prominent component in the composed vector.

In the present paper we offer a new approach to this formalization of the compositional meaning of adjectives and nouns that owes to both distributional VSMs and LVMs. Through this combination, we attempt to improve on earlier work in Almuhareb (2006) and Hartung and Frank (2010), which are both embedded in a purely distributional setting.

Specifically, we use Latent Dirichlet Allocation (LDA; Blei et al. (2003)) to train an attribute model that captures semantic information encoded in adjectives and nouns independently of one another. Following Hartung and Frank (2010), this model is embedded into a VSM that employs vector composition to combine the meaning of adjectives and nouns. We present two variants of LDA that differ in the way attributes are associated with the induced LDA topics: Controled LDA (C-LDA) and Labeled LDA (L-LDA; Ramage et al. (2009)). Both will be presented in detail in Section 3.

Our aims in this paper are two-fold: (i) We investigate LDA as a modeling framework in the attribute selection task, as its use of topics as latent variables may alleviate inherent sparsity problems faced by prior work using pattern-based (Almuhareb, 2006) or vector space models (Hartung and Frank, 2010). (ii) While these prior approaches were restricted to a confined set of 10 attributes, we will we apply our models on a much larger space of attributes, to probe their capacity on a more realistic data set.

The remainder of this paper is divided as follows. Section 2 reviews related work on distributional models of adjective semantics, and introduces the two frameworks in which we ground our approach: LVMs and VSMs. In Section 3 we introduce two LDA models for attribute selection: C-LDA and L-LDA. Section 4 describes the settings for two experiments: In the first experiment, we perform attribute selection confined to a space of 10 attributes to compare against prior work. In the second setting we perform attribute selection on a large scale, using 206 attributes. Section 5 presents and discusses the results. Section 6 concludes.

2 Related Work

Distributional models of adjective semantics. Almuhareb (2006) aims at capturing the relationship between adjectives and attributes based on lexico-syntactic patterns, such as *the ATTR of the * is ADJ*. Apart from inherent sparsity issues, his approach does not account for the compositional nature of the problem, as the contextual information contributed by a noun is neglected: For instance, his model is unable to predict that *hot* is unlikely to denote TASTE in the context of *summer*, other than in *hot meal*.

Compositionality of adjective-noun phrases and how it can be adequately modeled in VSMs is the main concern in Baroni and Zamparelli (2010) and Guevara (2010), who are in search of the best composition operator for combining adjective with noun meanings. While these works adhere to a purely latent representation of meaning, Hartung and Frank (2010) include attributes as symbolic 'hidden' meanings of adjectives, nouns and adjective-noun phrases in a distributional VSM.

Finally, a large body of work dealing with compositionality in distributional frameworks is not confined to the special case of adjective-noun composition (Mitchell and Lapata (2008), Rudolph and Giesbrecht (2010), i.a.). All these approaches regard composition as a process combining vectors (or matrices, resp.) to yield a new, contextualized vector representation within the same semantic space.

Latent Dirichlet Allocation, aka. Topic Models (TMs). LDA is a generative probabilistic model

[1] The figure is adopted from the distributional setting of Hartung and Frank (2010), with component values defined by pattern frequency counts for the chosen attribute nouns.

for document collections. Each document is represented as a mixture over latent *topics*, where each topic is a probability distribution over words (Blei et al., 2003). These topics can be used as dense features for, e.g., document clustering. Depending on the number of topics, which has to be pre-specified, the dimensionality of the document representation can be considerably reduced in comparison to simple bag-of-words models. The remainder of this paper will assume some familiarity with LDA and the LDA terminology as introduced in Blei et al. (2003).

Recent work investigates ways of accommodating supervision with LDA, e.g. supervised topic models (Blei and McAuliffe, 2007), Labeled LDA (L-LDA) (Ramage et al., 2009) or DiscLDA (Lacoste-Julien et al., 2008). We will discuss L-LDA in Section 3.

Distributional VSMs and TMs. The idea to integrate topic models and VSMs goes back to Mitchell and Lapata (2009) who build a distributional model with dimensions set to topics over bag-of-words features. In their setting, LDA merely serves the purpose of dimensionality reduction, whereas our particular motivation is to use topics as probabilistic indicators for the prediction of attributes as semantic target categories in adjective-noun composition. Mitchell and Lapata (2010) compare VSMs defined over bags of context words vs. latent topics in a similarity judgement task. Their results indicate that a multiplicative setting works best for vector composition in word-based models, while vector addition is better suited for topic vectors.

3 Topic Models for Attribute Selection

3.1 Using LDA for modeling lexical semantics

Recently, LDA has been used for problems in lexical semantics, where the primary goal is not document modeling but the induction of semantic knowledge from high-dimensional co-occurrence data. Ritter et al. (2010) and Ó Séaghdha (2010) model selectional restrictions of verbs by inducing topic distributions that characterize 'mixtures of topics' observed in verb argument positions. As a basis for LDA modeling, they collect *pseudo-documents*, i.e. bags of words that co-occur in syntactic argument positions.

We apply a similar idea to the attribute selection problem: we collect pseudo-documents that characterize attributes by adjectives and nouns that co-occur with the attribute nouns in local contextual relations. The topic distributions obtained from fitting an LDA model to the collection of these pseudo-documents can then be injected into semantic vector representations for adjectives and nouns.

In its original statement, LDA is a fully unsupervised process (apart from the desired number of topics which has to be specified in advance) that estimates topic distributions over documents θ_d and topic-word distributions ϕ_t with topics represented as latent variables. Estimating these parameters on a document collection yields *topic proportions* $P(t|d)$ and topic distributions $P(w|t)$ that can be used to compute a smooth distribution $P(w|d)$ as in (2), where t denotes a latent topic, w a word and d a document in the corpus.

$$P(w|d) = \sum_t P(w|t)P(t|d) \qquad (2)$$

Being designed for exploratory rather than discriminative analysis, LDA does not intend conditioning of words or topics on external categories. That is, the resulting topics cannot be related to previously defined target categories. For attribute selection, the LDA-inferred topics need to be linked to semantic attributes. Therefore, we apply two extensions of standard LDA that are capable of taking supervised category information into account, either implicitly or directly, by including an additional observable variable into the generative process.

In general, LVMs can be expected to overcome sparsity issues that are frequently encountered in distributional models. This positive smoothing effect is achieved by marginalization over the latent variables (cf. Prescher et al. (2000)). For instance, it is unlikely to observe a dependency path linking the adjective *mature* to the attribute MATURITY. Such a relation is more likely for *young*, for example. If *young* co-occurs with *mature* in a different pseudo-document (AGE might be a candidate), this results in a situation where (i) *young* and *mature* share one or more latent topics and (ii) the topic proportions for the attributes MATURITY and AGE will become similar to the extent of common words in their pseudo-documents. Consequently, the final attribute model is expected to assign a (small) positive probability to the relation between *mature* and MATURITY without observing it in the training data.

3.2 Controled LDA

The generative story behind C-LDA is equivalent to standard LDA. However, the collection of pseudo-documents used as input to C-LDA is structured in a controled way such that each document conveys semantic information that specifically characterizes the individual categories of interest (attributes, in our case). In line with the distributional hypothesis (Harris, 1968), we consider the pseudo-documents constructed in this way as distributional fingerprints of the meaning of the corresponding attribute.

The contents of the pseudo-documents are selected along syntactic dependency paths linking each attribute noun to meaningful context words (adjectives and nouns).[2] A corpus consisting of the two sentences in (3), e.g., yields a pseudo-document for the attribute noun SPEED containing *car* and *fast*.

(3) What is the speed of this car? The machine runs at a very fast speed.

Though we are ultimately interested in triples of attributes, adjectives and nouns that define the compositional semantics of adjective-noun phrases (cf. (1)), C-LDA is only exposed to binary tuples between attributes and adjectives or nouns, respectively. This is in line with Hartung and Frank (2010), who obtained substantial performance improvements by splitting the ternary relation into two binary relations.

Presenting LDA with pseudo-documents that characterize individual target attributes imports supervision into the LDA process in two respects: the estimated topic proportions $P(t|d)$ will be highly attribute-specific, and similarly so for the topic distributions $P(w|t)$. This makes the model more expressive for the ultimate labeling task. Moreover, since C-LDA collects pseudo-documents focused on individual target attributes, we are able to link external categories to the generative process by heuristically labeling pseudo-documents with their respective attribute as target category. Thus, we approximate $P(w|a)$, the probability of a word given an attribute, by $P(w|d)$ as obtained from LDA:

[2]The dependency paths, together with the set of attribute nouns of interest, have to be manually specified. See the supplementary material for the full list of dependency paths used.

1 For each topic $k \in \{1, \ldots, K\}$:
2 Generate $\beta_k = (\beta_{k,1}, \ldots, \beta_{k,V})^T \sim Dir(\cdot \mid \eta)$
3 For each document d:
4 For each topic $k \in \{1, \ldots, K\}$
5 Generate $\Lambda_k^{(d)} \in \{0, 1\} \sim Bernoulli(\cdot \mid \Phi_k)$
6 Generate $\alpha^{(d)} = L^{(d)} \times \alpha$
7 Generate $\theta^{(d)} = (\theta_{l_1}, \ldots, \theta_{l_{M_d}})^T \sim Dir(\cdot \mid \alpha^{(d)})$
8 For each i in $\{1, \ldots, N_d\}$:
9 Generate $z_i \in \{\lambda_1^{(d)}, \ldots, \lambda_{M_d}^{(d)}\} \sim Mult(\cdot \mid \theta^{(d)})$
10 Generate $w_i \in \{1, \ldots, V\} \sim Mult(\cdot \mid \beta_{z_i})$

Figure 2: L-LDA generative process (Ramage et al. 2009)

$$P(w|a) \approx P(w|d) = \sum_t P(w|t)P(t|d) \qquad (4)$$

3.3 Labeled LDA

L-LDA (Ramage et al., 2009) extends standard LDA to include supervision for specific target categories, yet in a different way: (i) The generative process includes a second observed variable, i.e. each document is explicitly labeled with a target category. A document may be labeled with an arbitrary number of categories; unlabeled documents are also possible. However, L-LDA permits only binary assignments of categories to documents; probabilistic weights over categories are not intended. (ii) Contrary to LDA, where the number of topics has to be specified in advance, L-LDA sets this parameter to the number of unique target categories. Moreover, the model is constrained such that documents may be assigned only those topics that correspond to their observable category label(s). That is, latent topics t in the standard formulation of LDA (2) are constrained to correspond to explicit labels a.

More specifically, L-LDA extends the generative process of LDA by constraining the topic distributions over documents $\theta^{(d)}$ to only those topics that correspond to the document's set of labels $\Lambda^{(d)}$. This is done by projecting the parameter vector of the Dirichlet topic prior α to a lower-dimensional vector $\alpha(d)$ whose topic dimensions correspond to the document labels.

This extension is integrated in steps 5 and 6 of Fig. 2: First, in step 5, the document's labels $\Lambda^{(d)}$ are generated for each topic k. The resulting vector of document's labels $\lambda^{(d)} = \{k \mid \Lambda_k^{(d)} = 1\}$ is used to define a document-specific label projection matrix

$L^{(d)}_{|\lambda^{(d)}|\times K}$, such that $L^{(d)}_{ij} = 1$ if $\lambda^{(d)}_i = j$, and 0 otherwise. This matrix is used in step 6 to project the Dirichlet topic prior α to a lower-dimensional vector $\alpha^{(d)}$, whose topic dimensions correspond to the document labels. Topic proportions are then, in step 7, generated for this reduced parameter space.

In our instantiation of L-LDA, we collect pseudo-documents for attributes exactly as for C-LDA. Documents are labeled with exactly one category, the attribute noun. Note that, even though the relationship between documents and topics is fixed, the one between topics and words is not. Any word occurring in more than one document will be assigned a non-zero probability for each corresponding topic.

Thus, with regard to attribute modeling, C-LDA and L-LDA build an interesting pair of opposites: The L-LDA model assumes that attributes are semantically primitive in the sense that they cannot be decomposed into smaller topical units, whereas words may be associated with several attributes at the same time. C-LDA, at the other end of the spectrum, licenses semantic variability on both the attribute and the word level. Particularly, a word might be associated with some of the topics underlying an attribute, but not with all of them, and an attribute can be characterized by multiple topics.

3.4 Vector Space Framework

For integrating the information obtained from C-LDA or L-LDA into a distributional VSM, we follow Hartung and Frank (2010): Adjectives and nouns are modeled as independent semantic vectors along their relationship to attributes; the most prominent attribute(s) that represent the hidden meaning of adjective-noun phrases are selected from their composition (cf. Fig. 1).

The dimensions of the VSM are set to the preselected attributes. Semantic vectors are computed for all adjectives and nouns occurring at least five times in the pseudo-documents. Vector component values $v_{\langle w,a\rangle}$ are derived from the C-LDA and L-LDA models in different ways: with C-LDA we obtain $P(w|a)$ by approximation from $P(w|d)$ (cf. equation (4)), while in L-LDA we obtain $P(w|a)$ directly from the induced topic-word distribution ϕ_t, through labeled topics $t = a$ (cf. equation (2)).

Vector composition is defined as *vector multiplication* ($\times$) or *vector addition* ($+$).

For attribute selection on the composed vector, we use two methods we found to perform best in Hartung and Frank (2010): Entropy Selection (ESel) and Most Prominent Component (MPC). ESel measures entropy over the vector components to identify components that encode a high amount of information. It selects all attributes that lead to an increase of entropy when suppressed from the vector representation. If no informative components can be detected in a vector due to a very broad, flat distribution of the probability mass (cf. $\vec{b}$ in Fig. 1), ESel yields an empty list. MPC always chooses exactly one vector component, i.e. the one with the highest value.

4 Experimental Settings

Attribute selection over small and large semantic spaces. We evaluate the performance of the VSMs based on C-LDA and L-LDA in two experimental settings, contrasting the problem of attribute selection on semantic spaces of radically different dimensionality, using sets of 10 vs. 206 attributes.

Evaluation measures. We evaluate against two gold standards consisting of adjective-noun phrases (or adjective-noun pairs) and their associated attribute meanings. We report precision, recall and f_1-score. Where appropriate, we test differences in the performance of various model configurations for statistical significance in a randomized permutation test (Yeh, 2000), using the `sigf` tool (Padó, 2006).

Baselines. We compare our models against two baselines, PATTVSM and DEPVSM. PATTSVM is reconstructed from Hartung and Frank (2010). It is grounded in a selection of lexical patterns that identify the target elements (adjectives and nouns) for the vector basis elements (i.e., the attribute nouns) in a local context window. The component values are defined using raw frequency counts over the extracted patterns. DEPVSM is similar to PATTVSM; however, it relies on dependency paths that connect the target elements and attributes in local contexts. The paths are identical to the ones used for constructing pseudo-documents in C-LDA and L-LDA. As in PATTVSM, the vector components are set to raw frequencies over extracted paths.

Implementations. To implement our models, we rely on MALLET (McCallum, 2002) for C-LDA and

the Stanford Topic Modeling Toolbox[3] for L-LDA. In both cases, we run 1000 iterations of Gibbs sampling, using default values for all hyperparameters.

Data set for attribute selection over 10 attributes. The first experiment is conducted on the data set used in Hartung and Frank (2010). It consists of 100 adjective-noun pairs manually annotated for ten attributes: COLOR, DIRECTION, DURATION, SHAPE, SIZE, SMELL, SPEED, TASTE, TEMPERATURE, WEIGHT. To enable comparison, the dimensions of our models are set to exactly these attributes.

Data set for attribute selection over a large semantic space (206 attributes). In the second experiment, we max out the attribute selection task to a much larger set of attributes in order to analyze the difficulty of the task on more representative data. We automatically construct a data set of adjective-noun phrases labeled with appropriate attributes from WordNet 3.0 (Fellbaum, 1998), relying on the assumption that examples given in glosses correspond to the respective word sense of the adjective. We first extract all adjectives that are linked to at least one attribute synset by the `attribute` relation. Next, we run the glosses of these adjectives (3592 in number) through TreeTagger (Schmid, 1994) to find examples of adjectives modifying nouns in attributive constructions. The resulting adjective-noun phrases are labeled with the attribute label linked to the given adjective sense.

This method yields 7901 labeled adjective-noun phrases. They are divided into development and test data according to a sampling procedure that respects the following criteria: (i) Both sets must contain all attributes with an equal number of phrases for each attribute; (ii) phrases with both elements contained in CoreWordNet[4] are preferred, while others are only considered if necessary to satisfy the first criterion. This procedure yields 496/345 phrases in the development/test set, distributed over 206 attributes[5].

[3] `http://nlp.stanford.edu/software/tmt/`.

[4] A subset of WordNet restricted to the 5000 most frequently used word senses. Available from: `http://wordnetcode.princeton.edu/standoff-files/core-wordnet.txt`

[5] If an attribute provides only one example, this was added to the development set. Therefore, the test set only comprises 206 attributes, while all models were trained on 262 attributes obtained from WordNet in the first extraction step.

Training data. The pseudo-documents are collected from dependency paths obtained from section 2 of the parsed pukWaC corpus (Baroni et al., 2009).

5 Discussion of Results

5.1 Experiment 1

In Experiment 1, we evaluate the performance of C-LDA and L-LDA on the attribute selection task over 10 attributes against the pattern-based and dependency-based models PATTVSM and DEPVSM as competitive baselines. Besides a comparison to standard VSMs, we are especially interested in the relative performance of the LDA models. Given that C-LDA and L-LDA estimate attribute-specific topic distributions in the structured pseudo-documents under different assumptions regarding the correspondence of attributes and topics (cf. Sec. 3.2 and 3.3), we expect the two LDA variants to differ in their capability to capture the topic distributions in the labeled pseudo-documents.

5.1.1 Attribute Selection for 10 Attributes

Tables 1 and 2 summarize the results for attribute selection over 10 attributes against the labeled adjective-noun pairs in the test set, using ESel and MPC as selection functions on vectors composed by multiplication (Table 1) and addition (Table 2). The results reported for C-LDA correspond to the best performing model (with number of topics set to 42, as this setting yields the best and most constant results over both composition operators).

C-LDA shows highest f-scores and recall over all settings, and highest precision with vector addition.[6] In line with Mitchell and Lapata (2010) (cf. Sec. 2), we obtain the best overall results with vector addition (ESel: P: 0.55, R: 0.66, F: 0.61; MPC: P: 0.59, R: 0.71, F: 0.64). The difference between C-LDA and L-LDA is small but significant for vector multiplication; for vector addition, it is not significant.

Compared to the LDA models, the VSM baselines

[6] In Tables 1 and 2, statistical significance of the differences between the models is marked by the superscripts L, D and P, denoting a significant difference over **L**-LDA, **D**epVSM and **P**attVSM, respectively. All differences reported are significant at $p < 0.05$, except for the difference between C-LDA and L-LDA in Table 3 ($p < 0.1$).

	ESel			MPC		
	P	R	F	P	R	F
C-LDA	0.58	**0.65**	**0.61**^{L,P}	0.57	**0.64**	**0.60**
L-LDA	**0.68**	0.54	0.60^{D}	0.55	0.61	0.58^{D}
DepVSM	0.48	0.58	0.53^{P}	0.57	0.60	0.58
PattVSM	0.63	0.46	0.54	**0.60**	0.58	0.59

Table 1: Attribute selection over 10 attributes ($\times$)

	ESel			MPC		
	P	R	F	P	R	F
C-LDA	**0.55**	**0.66**	**0.61**^{D,P}	**0.59**	**0.71**	**0.64**
L-LDA	0.53	0.57	0.55^{D,P}	0.50	0.45	0.47^{D,P}
DepVSM	0.38	0.65	0.48^{P}	0.57	0.60	0.58
PattVSM	0.71	0.35	0.47	0.47	0.56	0.51

Table 2: Attribute selection over 10 attributes ($+$)

are competitive, but tend to perform lower. This effect is statistically significant for ESel with vector multiplication: each of the LDA models statistically significantly outperforms one of the VSM models, DEPVSM and PATTVSM. With ESel and vector addition, both LDA models outperform both VSM models statistically significantly. The LDA$_{ESel,+}$ models outperform the PATTVSM$_{ESel,+}$ model of Hartung and Frank (2010) by a high margin in f-score: +0.14 for C-LDA; +0.08 for L-LDA. Compared to the stronger multiplicative settings PATTVSM$_{ESel,\times}$ and PATTVSM$_{MPC,\times}$ this still represents a plus of +0.07 and +0.02 in f-score, respectively. We further observe a clear improvement of the LDA models over the VSM models in terms of recall (+0.20, C-LDA$_{ESel,+}$ vs. PATTVSM$_{ESel,\times}$), at the expense of some loss in precision (-0.08, C-LDA$_{ESel,+}$ vs. PATTVSM$_{ESel,\times}$). This clearly confirms a stronger generalization power of LDA compared to VSM models.

With regard to selection functions, we observe that MPC tends to perform better for the VSM models, while ESel is more suitable in the LDA models.

Figures 3 and 4 display the overall performance curve ranging over different topic numbers for C-LDA$_{ESel,+}$ and C-LDA$_{ESel,\times}$ – compared to the remaining models that are not dependent on topic size. For topic numbers smaller than the attribute set size, C-LDA underperforms, for obvious reasons. Increasing ranges of topic numbers to 60 does not show a linear effect on performance. Parameter settings with performance drops below the VSM baselines are rare, which holds particularly for vector addition at topic ranges larger than 10. With vector addition, C-LDA outperforms L-LDA in almost all configurations, yet at an overall lower performance level of L-LDA (0.55 with addition vs. 0.6 with multiplication). Note that in the multiplicative setting, C-LDA reaches the performance of L-LDA only in its best configurations, while with vector addition it obtains high performance that exceeds L-LDA's top f-score of 0.6 for topic ranges between 10 and 20.

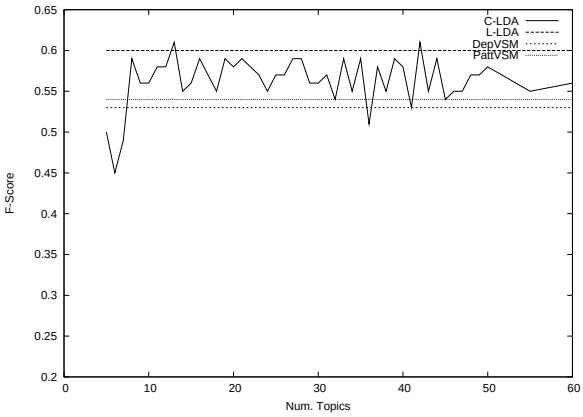

Figure 3: Performance of C-LDA$_{ESel,\times}$ for different topic numbers, compared against all other models

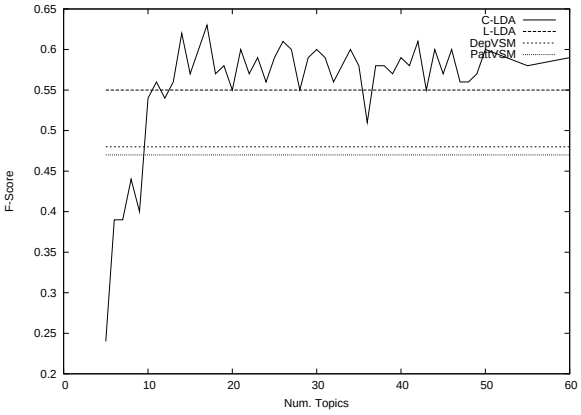

Figure 4: Performance of C-LDA$_{ESel,+}$ for different topic numbers, compared against all other models

Based on these observations, vector addition seems to offer the more robust setting for C-LDA, the model that is less strict with regard to topic-attribute correspondences. Vector multiplication, on the other hand, is more suitable for L-LDA and its stricter association of topics with class labels.

5.1.2 Smoothing Power of LDA Models

Our hypothesis was that LDA models should be better suited for dealing with sparse data, compared

	ESel			MPC		
	P	R	F	P	R	F
C-LDA	**0.39**	**0.31**	**0.35**	**0.37**	**0.27**	**0.32**
L-LDA	0.30	0.18	0.23	0.20	0.18	0.19
DepVSM	0.20	0.10	0.13	0.37	0.26	0.30
PattVSM	0.00	0.00	0.00	0.00	0.00	0.00

Table 3: Performance figures on sparse vectors ($\times$)

	ESel			MPC		
	P	R	F	P	R	F
C-LDA	**0.43**	**0.33**	**0.38**	**0.44**	**0.28**	**0.34**
L-LDA	0.34	0.16	0.22	0.37	0.18	0.24
DepVSM	0.16	0.17	0.17	0.36	0.21	0.27
PattVSM	0.13	0.04	0.06	0.17	0.25	0.20

Table 4: Performance figures on sparse vectors ($+$)

to pattern-based or purely distributional approaches. While this is broadly confirmed in the above results by global gains in recall, we conduct a special evaluation focused on those pairs in the test set that suffer from sparse data. We selected all adjective and noun vectors that did not yield any positive component values in the PATTVSM model. The 22 adjective-noun pairs in the test set affected by these 'zero vectors' were evaluated using the remaining models.

The results in Tables 3 and 4 yield a very clear picture: C-LDA obtains highest precision, recall and f-score across all settings, followed by L-LDA and DEPVSM$_{ESel}$, while their ranks are reversed when using MPC. Again, MPC works better for the VSM models, ESel for the LDA models. Vector addition performs best for C-LDA with f-scores of 0.38 and 0.34 – outperforming the pattern-based results on sparse vectors by orders of magnitude.

5.2 Experiment 2

Experiment 2 is designed to max out the space of attributes to be modeled, to assess the capacity of both LDA models and the DEPVSM baseline model in the attribute selection task on a large attribute space.[7] In contrast to Experiment 1, with its confined semantic space of 10 target attributes, this represents a huge undertaking.

5.2.1 Large-scale Attribute Selection

Table 5 (column **all**) displays the performance of all models on attribute selection over a range of 206 dimensions, contrasting vector addition and multiplication. The number of topics was set to 400. As the overall performance is close to 0 for both composition methods, no parameter setting can be identified as particularly suited for this large-scale attribute selection task. The differences between the three models are very small and not significant[8].

[7] We did not apply PATTVSM to this large-scale experiment, as only poor performance can be expected.

	all		property	
	$\times$	$+$	$\times$	$+$
C-LDA	0.04	0.02	$0.18^{L,D}$	0.10^{D}
L-LDA	0.03	0.04	0.15	0.15
DepVSM	0.02	0.02	0.12	0.07

Table 5: Performance figures (in f-score) of C-LDA$_{ESel}$ on 206 (all) and 73 property attributes (property)

	all			property		
	P	R	F	P	R	F
WIDTH	0.67	1.00	0.80	1.00	0.50	0.67
WEIGHT	0.80	0.57	0.67	0.50	0.57	0.53
MAGNETISM	0.50	1.00	0.67			
SPEED	0.50	0.50	0.50	1.00	0.50	0.67
TEXTURE	0.33	1.00	0.50	0.33	1.00	0.50
DURATION	0.50	0.50	0.50	1.00	1.00	1.00
TEMPERATURE	0.30	0.75	0.43	0.43	0.75	0.55
AGE	0.33	0.50	0.40			
THICKNESS	1.00	0.25	0.40	0.50	0.13	0.20
DEGREE	1.00	0.20	0.33			
LENGTH	0.17	1.00	0.29	0.50	1.00	0.67
DEPTH	1.00	0.14	0.25	1.00	0.86	0.92
ACTION	0.17	0.50	0.25			
LIGHT	0.33	0.17	0.22	0.20	0.17	0.18
POSITION	0.14	0.25	0.18	0.20	0.25	0.22
SHARPNESS				1.00	1.00	1.00
SERIOUSNESS				0.50	1.00	0.67
COLOR	0.13	0.25	0.17	0.29	0.50	0.36
LOYALTY				1.00	1.00	1.00
average	0.49	0.54	0.51	0.63	0.63	0.63

Table 6: Attribute selection on 206 attributes (all) and 73 property attributes (property); performance figures of C-LDA$_{ESel,\times}$ for best attributes (F>0)

5.2.2 Focused Evaluation and Data Analysis

To gain a deeper insight into the modeling capacity of the LDA models for this large-scale selection task, Table 6 (column **all**) presents a partial evaluation of attributes that could be assigned to adjective-noun pairs with an f-score >0 by C-LDA$_{ESel,\times}$.

Despite the disappointing overall performance of

[8] Again, statistically significant differences are marked by superscripts (cf. footnote 6). All differences reported are significant at $\alpha < 0.05$.

	prediction	correct
thin layer	THICKNESS	THICKNESS
heavy load	WEIGHT	WEIGHT
shallow water	DEPTH	DEPTH
short holiday	DURATION	DURATION
attractive force	MAGNETISM	MAGNETISM
short hair	LENGTH	LENGTH
serious book	DIFFICULTY	MIND
blue line	COLOR	UNION
weak president	POSITION	POWER
fluid society	REPUTE	CHANGEABLENESS
short flight	DISTANCE	DURATION
rough bark	TEXTURE	EVENNESS
faint heart	CONSTANCY	COWARDICE

Table 7: Sample of correct and false predictions of C-LDA$_{ESel,\times}$ in Experiment 2

the LDA models on this large attribute space, it is remarkable that C-LDA is able to induce distinctive topic distributions for a number of attributes with up to 0.51 f-score with balanced precision and recall, a moderate drop of only -0.10 relative to the corresponding model induced over 10 attributes.

Raising the attribute selection task from 10 to 206 attributes poses a true challenge to our models, by the sheer size and diversity of the semantic space considered. Table 7 gives an insight into the nature of the data and the difficulty of the task, by listing correct and false preditions of C-LDA for a small sample of adjective-noun pairs. Possible explanations for false predictions are manifold, among them near misses (e.g. *serious book*, *weak president*, *short flight*, *rough bark*), idiomatic expressions (e.g. *faint heart*, *blue line*) or questionable labels provided by WordNet (e.g. *serious book*).

As seen above, C-LDA achieves relatively high performance figures on selected attributes (cf. Table 6, col. **all**). In order to identify what makes these attributes different from others that resist successful modeling, we investigated three factors: (i) the amount of training data available for each attribute, (ii) the ambiguity rate per attribute, and (iii) their ontological subtype.

(i) Measuring the dependence between training data size and f-score per attribute shows that a large amount of training data is generally helpful, but not the decisive factor (Pearson's $r = 0.19$, $p < 0.01$).

(ii) The ambiguity rate AR_{attr} per attribute $attr$ is computed by averaging over all test pairs TP_{attr} labeled with $attr$, counting the total number of attributes $attr'$ that are associated with each adjective in pairs $\langle adj, n\rangle \in TP_{attr}$ in WordNet:

$$AR_{attr} = \frac{\sum_{attr'} \sum_{\langle adj,n\rangle \in TP_{attr}} |\langle adj, attr'\rangle_{WN}|}{|TP_{attr}|}$$

Correlating this figure with the performance per attribute in terms of f-score yields only a small positive correlation (Pearson's $r = 0.23$, $p < 0.01$). In fact, the qualitative analysis in Table 7 shows that C-LDA is capable of assigning meaningful attributes to adjective-noun phrases not only in easy, but also ambiguous cases (cf. *shallow water*, where DEPTH is the only attribute provided for *shallow* in WordNet vs. *short holiday*, *short hair* or *short flight*).

(iii) Although the 206 attributes used in Exp. 2 are rather diverse, including concepts such as HEIGHT, KINDNESS or INDIVIDUALITY, we observe a high number of attributes from Exp. 1 that are successfully modeled in Exp. 2 (5 out of 10, cf. column **all** in Table 6). Given that they are categorized into the *property* class in WordNet[9], we presume that the varying performance across attributes might be influenced by their ontological subtype. This hypothesis is validated in a replication of Exp. 2, with training data limited to the 73 attributes pertaining to the *property* subtype in WordNet. The test set was restricted accordingly, resulting in 112 pairs that are linked to a *property* attribute.

The overall performance of the models in this experiment is shown in Table 5 (column **property**): With vector multiplication, the best-performing operation across all models, all models benefit considerably (+0.10 or more). C-LDA shows the largest improvement, significantly outperforming both L-LDA and DEPVSM. With vector addition, the performance gains are slightly lower in general. In this setting, L-LDA shows higher f-score than C-LDA, though this difference is not statistically significant. Still, C-LDA significantly outranges DEPVSM. Note that we can not show a significant difference between C-LDA$_{ESel,\times}$ and L-LDA$_{ESel,+}$, so the comparison between these models remains inconclusive here. Note further that the affinity of C-LDA with vector addition and L-LDA with vector multiplication, respectively, is inverted in the large-scale experiment (cf. Table 5).

[9]WordNet separates attributes into *properties*, *qualities* and *states*, among several others.

While these overall results are far from satisfactory, they still clearly indicate that the LDA models work effectively for at least a subset of attributes, and outperform the VSM baseline.

Again, a more detailed analysis is given in Table 6 (column **property**), showing the performance of the best individual property attributes ($F>0$) in the restricted experiment. Average performance of the best property attributes with $F>0$, individually, amounts to F=0.63[10]. In comparison to the unrestricted setting (cf. column **all**), nearly all property attributes benefit from model training on selective data. Exceptions are WIDTH, WEIGHT, THICKNESS, AGE, DEGREE and LIGHT. Thus, apparently, some of the adjectives associated with non-property attributes in the full set provide some discriminative power that is helpful to distinguish property types.

In a qualitative analysis of the 133 non-property attributes filtered out in this experiment, we find that the WordNet-SUMO mapping (Niles, 2003) does not provide differentiating definitions for about 60% of these attributes, linking them instead to a single *subjective assessment attribute*. This suggests that in many cases the distinctions drawn by WordNet are too subtle even for humans to reproduce.

6 Conclusion

This paper explored the use of LDA topic models in a semantic labeling task that predicts attributes as 'hidden' meanings in the compositional semantics of adjective-noun phrases. LDA topic models are expected to alleviate sparsity problems of distributional VSMs as encountered in prior work, by incorporating latent semantic information about attribute nouns. We investigated two variants of LDA that employ different degrees of supervision for associating topics with attributes.

Our contributions are as follows. We proposed two LDA models for the attribute selection task that import supervision for a target category parameter in different ways: L-LDA (Ramage et al., 2009) embeds the target categories into the LDA process, by defining a 1:1 correspondence of topics and target categories. C-LDA, by contrast, does not affect the LDA generative process. Here, we heuristically equate pseudo-documents with target categories, to approximate category-specific word-topic distributions. By adhering to standard LDA, C-LDA accommodates a greater variety in the distributions of topics to attribute-specific documents and words, as compared to L-LDA. Combining standard LDA topic modeling with a means of interpreting the induced topics relative to a set of external categories, C-LDA offers greater flexibility and expressiveness.

Our experimental results show that modeling attributes as latent or explicit topics with C-LDA and L-LDA, respectively, outperforms the purely distributional baseline model DEPVSM and PATTVSM of prior work. Targeted evaluation on sparse data points confirms that LDA models help to overcome inherent sparsity effects of VSMs. C-LDA and L-LDA are close in performance in Experiment 1. C-LDA outperforms L-LDA only with optimal topic parameter settings.

Finally, we probed the modeling capacity of LDA and VSM models on a vast space of 206 attributes. This task proved to be extremely difficult. However, we obtain respectable results on a subset of attributes denoting properties, where C-LDA performs best in quantitative performance measures. It yields highest f-scores in full and partial evaluation – both with the full-size attribute model, and when training and testing is restricted to property attributes. The differences are small, but statistically significant between the LDA models and the VSM baseline in a setting restricted to property attributes.

Data analysis indicates that our models perform more robustly on concrete attributes in contrast to abstract attribute types that lack clear categorization. This suggests that our approach to attribute selection is most appropriate for detecting attributes that reflect clear ontological distinctions.

However, there is ample space for improvement. In Hartung and Frank (2011), we show that the quality of the noun vectors lags behind the adjective vectors. This clearly affects the performance of our models in cases where the semantic contribution of the noun is decisive for disambiguation. Future work will focus on ways to enhance the noun vector representations through additional contextual features, to make them denser and more articulated in structure.

[10] In comparison, L-LDA$_{ESel,\times}$ yields an average f-score of 0.47 for attributes with $F>0$ in the property setting.

References

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

Marco Baroni and Alessandro Lenci. 2010. Distributional Memory. A General Framework for Corpus-based Semantics. *Computational Linguistics*, 36:673–721.

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*, East Stroudsburg, PA, pages 1183–1193.

M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43:209–226.

D. Blei and J. McAuliffe. 2007. Supervised topic models. *Neural Information Processing Systems*, 21.

David M. Blei, Andrew Ng, and Michael Jordan. 2003. Latent dirichlet allocation. *JMLR*, 3:993–1022.

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

Emiliano Guevara. 2010. A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, Stroudsburg, PA. Association for Computational Linguistics.

Zellig Harris. 1968. *Mathematical Structures of Language*. Wiley.

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, August.

Matthias Hartung and Anette Frank. 2011. Assessing interpretable, attribute-related meaning representations for adjective-noun phrases in a similarity prediction task. In *Proceedings of GEometrical Models of Natural Language Semantics (GEMS-2011)*, Edinburgh, UK.

Simon Lacoste-Julien, Fei Sha, and Michael I. Jordan. 2008. DiscLDA: Discriminative Learning for Dimensionality Reduction and Classification. In *NIPS*, volume 22.

Linlin Li, Benjamin Roth, and Caroline Sporleder. 2010. Topic models for word sense disambiguation and token-based idiom detection. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1138–1147, Uppsala, Sweden.

Andrew Kachites McCallum. 2002. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu.

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

Jeff Mitchell 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.

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

Ian Niles. 2003. Mapping WordNet to the SUMO Ontology. In *Proceedings of the IEEE International Knowledge Engineering conference*, pages 23–26, June.

Diarmuid Ó Séaghdha. 2010. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 435–444, Uppsala, Sweden, July. Association for Computational Linguistics.

Sebastian Padó, 2006. *User's guide to `sigf`: Significance testing by approximate randomisation*.

D. Prescher, S. Riezler, and M. Rooth. 2000. Using a probabilistic class-based lexicon for lexical ambiguity resolution. In *Proceedings of the 18th COLING*, pages 649–655.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, August 2009, pages 248–256.

Alan Ritter, Mausam, and Oren Etzioni. 2010. A Latent Dirichlet Allocation Method for Selectional Preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Uppsala, Sweden, July. Association for Computational Linguistics.

Sebastian Rudolph and Eugenie Giesbrecht. 2010. Compositional matrix-space models of language. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 907–916. Association for Computational Linguistics, July.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP). Manchester, U.K., 14–16 September 1994*.

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

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the Fourth Conference on Computational Language Learning (CoNLL-2000) and the Second Learning Language in Logic Workshop*, Lisbon, Portugal, pages 947–953.