Assessing Interpretable, Attribute-related Meaning Representations for Adjective-Noun Phrases in a Similarity Prediction Task

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

Abstract

We present a distributional vector space model that incorporates Latent Dirichlet Allocation in order to capture the semantic relation holding between adjectives and nouns along interpretable dimensions of meaning: The meaning of adjective-noun phrases is characterized in terms of ontological attributes that are prominent in their compositional semantics. The model is evaluated in a similarity prediction task based on paired adjective-noun phrases from the Mitchell and Lapata (2010) benchmark data. Comparing our model against a high-dimensional latent word space, we observe qualitative differences that shed light on different aspects of similarity conveyed by both models and suggest integrating their complementary strengths.

1 Introduction

This paper offers a comparative evaluation of two types of accounts to the compositional meaning of adjective-noun phrases. This comparison is embedded in a similarity judgement task that determines the semantic similarity of pairs of adjective-noun phrases. All models we consider establish the similarity of adjective-noun pairs by measuring similarity between vectors representing the meaning of the individual adjective-noun phrases. However, the models we investigate differ in the type of interpretation they assign to adjectives, nouns and the phrases composed from them.

One type of approach is represented by the classical vector space model (VSM) of Mitchell and Lapata (2010; henceforth: M&L). It represents the semantics of adjective-noun phrases in *latent semantic space*, based on dimensions defined by bags of context words. This classical model will be compared against a compositional analysis of adjective-noun phrases that represents adjectives and nouns along *interpretable dimensions* of meaning, i.e. discrete ontological attributes such as SIZE, COLOR, SPEED, WEIGHT. Here, lexical vectors for adjectives and nouns define possible attribute meanings as component values; vector composition is intended to elicit those attributes that are prominent in the meaning of the whole phrase. For instance, a composed vector representation of the phrase *hot pepper* is expected to yield high component values on the dimensions TASTE and SMELL, rather than TEMPERATURE. The underlying relations between adjectives and nouns, respectively, and the attributes they denote is captured by way of latent semantic information obtained from Latent Dirichlet Allocation (LDA; Blei et al. (2003)). Thus, we treat attributes as an abstract meaning layer that generalizes over latent topics inferred by LDA and utilize this interpretable layer as the dimensions of our VSM.

This approach has been shown to be effective in an *attribute selection* task (Hartung and Frank, 2011), where the goal is to predict the most prominent attribute(s) "hidden" in the compositional semantics of adjective-noun phrases. In this paper, our main interest is to assess the potential of modeling adjective semantics in terms of discrete, interpretable attribute meanings in a similarity judgement task, as opposed to a representation in latent semantic space that is usually applied to tasks of this kind.

Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics, EMNLP 2011, pages 52–61,
Edinburgh, Scotland, UK, July 31, 2011. ©2011 Association for Computational Linguistics

For this purpose, we rely on the evaluation data set of M&L which serves as a shared benchmark in the GEMS 2011 workshop. Their similarity judgement task, being tailored to measuring latent similarity, represents a true challenge for an analysis focused on discrete ontological attributes.

Our results show that the latent semantic model of M&L cannot be beaten by an interpreted analysis based on LDA topic models. However, we show substantial performance improvements of the interpreted analysis in specific settings with adapted training and test sets that enable focused comparison. An interesting outcome of our investigations is that – using an interpreted LDA analysis of adjective-noun phrases – we uncover divergences in the notions of similarity underlying the judgement task that go virtually unnoticed in a latent semantic VSM, while they need to be clearly distinguished in models focused on interpretable representations.

The paper is structured as follows: After a brief summarization of related work, Section 3 introduces *Controled LDA*, a weakly supervised extension to standard LDA, and explains how it can be utilized to inject interpretable meaning dimensions into VSMs. In Section 4, we describe the parameters and experimental settings for comparing our model to M&L's word-based latent VSM in a similarity prediction task. Section 5 presents the results of this experiment, followed by a thorough qualitative analysis of the specific strengths and weaknesses of both models in Section 6. Section 7 concludes.

2 Related Work

Recent work in distributional semantics has engendered different perspectives on how to characterize the semantics of adjectives and adjective-noun phrases.

Almuhareb (2006) aims at capturing the semantics of adjectives in terms of attributes they denote using lexico-syntactic patterns. His approach suffers from severe sparsity problems and does not account for the compositional nature of adjective-noun phrases, as it disregards the meaning contributed by the noun. It is therefore unable to perform disambiguation of adjectives in the context of a noun.

Baroni and Zamparelli (2010) and Guevara (2010) focus on how best to represent compositionality in adjective-noun phrases considering different types of composition operators. These works adhere to a fully latent representation of meaning, whereas Hartung and Frank (2010) assign symbolic attribute meanings to adjectives, nouns and composed phrases by incorporating attributes as dimensions in a compositional VSM. By holding the attribute meaning of adjectives and nouns in distinct vector representations and combining them through vector composition, their approach improves on both weaknesses of Almuhareb's work. However, their account is still closely tied to Almuhareb's pattern-based approach in that counts of co-occurrence patterns linking adjectives and nouns to attributes are used to populate the vector representations. These, however, are inherently sparse. The resulting model therefore still suffers from sparsity of co-occurrence data.

Finally, Latent Dirichlet Allocation, originally designed for tasks such as text classification and document modeling (Blei et al., 2003), found its way into lexical semantics. Ritter et al. (2010) and Ó Séaghdha (2010), e.g., model selectional restrictions of verb arguments by inducing topic distributions that characterize mixtures of topics observed in verb argument positions. Mitchell and Lapata (2009, 2010) were the first to use LDA-inferred topics as dimensions in VSMs.

Hartung and Frank (2011) adopt a similar approach, by embedding LDA into a VSM for adjective-noun meaning composition, with LDA topics providing latent variables for attribute meanings. That is, contrary to M&L, LDA is used to convey information about interpretable semantic attributes rather than latent topics. In fact, Hartung and Frank (2011) are able to show that "injecting" topic distributions inferred from LDA into a VSM alleviates sparsity problems that persisted with the pattern-based VSM of Hartung and Frank (2010).

Baroni et al. (2010) highlight two strengths of VSMs that incorporate interpretable dimensions of meaning: cognitive plausibility and effectiveness in concept categorization tasks. In their model, concepts are characterized in terms of salient properties and relations (e.g., *children* have *parents*, *grass* is *green*). However, their approach concentrates on nouns. Open questions are (i) whether it can be extended to further word classes, and (ii) whether the interpreted meaning layers are interoperable across word classes, to cope with compositionality. The present paper extends their work by offering a test case for an interpretable, compositional VSM, applied to adjective-noun composition with attributes as a shared meaning layer. Moreover, to our knowledge, we are the first to expose such a model to a pairwise similarity judgement task.

3 Attribute Modeling based on LDA

3.1 Controled LDA

This section introduces *Controled LDA* (C-LDA), a weakly supervised variant of LDA. We use C-LDA to model attribute information that pertains to adjectives and nouns individually. This information is “injected” into a vector-space framework as a basis for computing the attributes that are prominent in compositional adjective-noun phrases.

In its original statement, LDA is a fully unsupervised process that estimates topic distributions over documents θ_d and word-topic distributions ϕ_t with topics represented as hidden 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 (1), where t denotes a latent topic, w a word and d a document in the corpus.

$$P(w|d) = \sum_t P(w|t)P(t|d) \qquad (1)$$

While the generative story underlying both models is identical, C-LDA extends standard LDA by “implicitly” taking supervised category information into account. This allows for linking latent topics to interpretable semantic attributes. The idea is to collect *pseudo-documents* in a controlled way such that each document conveys semantic information about one specific attribute. The pseudo-documents are selected along syntactic dependency paths linking the respective attribute noun to meaningful context words (adjectives and nouns). A corpus consisting of the two sentences in (2), e.g., yields a pseudo-document for the attribute noun SPEED containing *car* and *fast*.

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

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

3.2 Embedding C-LDA into a VSM

The main difference of C-LDA compared to standard LDA is that the estimated topic proportions $P(t|d)$ of the former will be highly attribute-specific, and similarly so for the topic distributions $P(w|t)$. We experiment with two variants of VSMs that differ in the way they integrate attribute information inferred from C-LDA, denoted as C-LDA-A and C-LDA-T.

In C-LDA-A, the dimensions of the space are interpretable attributes. The vector components relating a target word w to an attribute a are set to $P(w|a)$. This probability is obtained from C-LDA by constructing the pseudo-documents as distributional fingerprints of the respective attribute, as described in Section 3.1 above:

$$P(w|a) \approx P(w|d) = \sum_t P(w|t)P(t|d) \qquad (3)$$

C-LDA-T capitalizes on latent topics as dimensions; the vector components are set to the topic proportions $P(w|t)$ as directly obtained from C-LDA.[1]

4 Parameters and Experimental Settings

Data. Our experiments are based on the adjective-noun section of M&L’s 2010 evaluation data set[2]. It consists of 108 pairs of adjective-noun phrases that were rated for similarity by human judges.

[1] The “topics as dimensions” approach has also been used by Mitchell and Lapata (2010) for dimensionality reduction. In their word space model, however, this setting leads to a decrease in performance on adjective-noun phrases. Therefore, we do not compare ourselves to this instantiation of their model in this paper.

[2] Available from: `http://homepages.inf.ed.ac.uk/s0453356/share`

Models. We contrast the two LDA-based models (i, ii) C-LDA-A and C-LDA-T with two standard VSMs: (iii) a re-implementation of the latent VSM of M&L and (iv) a dependency-based VSM (DepVSM) which relies on dependency paths that connect the target elements and attribute nouns in local contexts. The paths are identical to the ones used for constructing pseudo-documents in (i) and (ii). Thus, DepVSM relies on the same information as C-LDA-A and C-LDA-T, without capitalizing on the smoothing power provided by LDA.

In the C-LDA models, we experiment with several topic number settings. Depending on the number of attributes $|A|$ contained in the training material (see below), we train one model instance for each topic number in the range from $0.5 \cdot |A|$ to $2 \cdot |A|$. For our LDA implementations, we use MALLET (McCallum, 2002). We run 1000 iterations of Gibbs sampling with hyperparameters set to the default values.

Training data. For C-LDA-A, C-LDA-T and DepVSM we apply two different training scenarios: In the first setting, we collect pseudo-documents instantiating 262 attribute nouns that are linked to adjectives by an `attribute` relation in WordNet (Fellbaum, 1998). The topic distributions induced from this data cover the broadest space of attribute meanings we could produce from WordNet[3]. In a second setting, we assume the presence of an "oracle" that confines the training data to a subset of 33 attribute nouns that are linked to those adjectives that actually occur in the M&L test set, to allow for a focused evaluation. In both C-LDA variants, all adjectives and nouns occurring at least five times in the pseudo-documents become target elements in the VSM. The pseudo-documents are collected along dependency paths extracted from section 2 of the pukWaC corpus (Baroni et al., 2009). The same settings are used for training the DepVSM model.

As the M&L model is not intended to reflect attribute meaning, the training data for this model remains constant. Like M&L, we set the target elements of this model to all types contained in the complete evaluation data set (including nouns, adjectives and verbs) and select the 2000 context words that co-occur most frequently with these targets in pukWaC_2 as the dimensions of the space.

Filters on test set. Given the different types of "semantic gist" of the models described above, we expect that the LDA models perform best on those test pairs that involve attributes known to the model. To test this expectation, we compile a restricted test set containing 43 pairs $(adj_1\ n_1,\ adj_2\ n_2)$ where both adj_1 and adj_2 bear an attribute meaning according to WordNet.

Composition operators. In our experiments, we use a subset of the operators proposed by Mitchell and Lapata (2010) to obtain a compositional representation of adjective-noun phrases from individual vectors: vector multiplication ($\times$; best operator in M&L's experiments on adjective-noun phrases) and vector addition ($+$). Besides, in order to assess the contribution of individual vectors in the composition process, we experiment with two "composition surrogates" by taking the individual adjective (ADJ-only) or noun vector (N-only) as the result of the composition process.

Evaluating the models. The models described above are evaluated against the human similarity judgements data provided by Mitchell and Lapata (2010) as follows: We compute the cosine similarity between the composed vectors representing the adjective-noun phrases in each test pair. Next, we measure the correlation between the model scores and the human judgements in terms of Spearman's ρ, where each human rating is treated as an individual data point. The correlation coefficient finally reported is the average over all instances[4] of one model. For completeness, we also report the correlation score of the best model instance and the standard deviation over all model instances.

5 Discussion of Results

Results on complete test set. Table 1 displays the results achieved by the VSMs based on C-LDA and

[3] Note that in Hartung and Frank (2011) only a subset of these attributes, mainly those characterized as *properties* in WordNet, could be successfully modeled, at overall moderate performance levels.

[4] In fact, only those model instances resulting in a significant correlation with the human judgements ($p < 0.05$) are taken into account. This way, we eliminate both inefficient and overly optimistic model instances.

		+			×			ADJ-only			N-only		
		avg	best	σ	avg	best	σ	avg	best	σ	avg	best	σ
262 attrs	C-LDA-A	0.19	0.25	0.05	0.15	0.20	0.04	0.17	0.23	0.04	0.11	0.23	0.06
	C-LDA-T	0.19	0.24	0.02	**0.28**	0.31	0.02	0.20	0.24	0.02	0.18	0.24	0.03
	M&L	0.21			**0.34**			0.19			**0.27**		
	DepVSM	-0.09			-0.09			-0.14			-0.08		
33 attrs	C-LDA-A	**0.23**	0.27	0.02	0.21	0.24	0.01	**0.27**	0.29	0.01	0.17	0.22	0.02
	C-LDA-T	0.21	0.28	0.03	0.14	0.23	0.04	0.22	0.27	0.03	0.10	0.21	0.06
	M&L	0.21			**0.34**			0.19			**0.27**		
	DepVSM	0.21			0.20			**0.27**			0.19		

Table 1: Correlation coefficients (Spearman's ρ) for different training sets, complete test set

		+			×			ADJ-only			N-only		
		avg	best	σ	avg	best	σ	avg	best	σ	avg	best	σ
262 attrs (filtered)	C-LDA-A	0.22	0.31	0.07	0.12	0.30	0.11	0.18	0.30	0.08	0.17	0.28	0.07
	C-LDA-T	0.25	0.30	0.03	0.26	0.35	0.04	0.24	0.29	0.04	0.19	0.23	0.04
	M&L	**0.38**			**0.40**			0.24			**0.43**		
	DepVSM	0.08			-0.09			0.06			-0.07		
33 attrs (filtered)	C-LDA-A	0.29	0.32	0.02	0.31	0.36	0.02	0.34	0.38	0.02	0.09	0.18	0.04
	C-LDA-T	0.26	0.36	0.05	0.14	0.30	0.09	0.28	0.38	0.07	0.03	0.18	0.08
	M&L	**0.38**			**0.40**			0.24			**0.43**		
	DepVSM	0.34			0.32			**0.35**			0.19		

Table 2: Correlation coefficients (Spearman's ρ) for different training sets and filtered test sets

the M&L word space model on the full adjective-noun test set. The table is split into an upper and a lower part containing the different results obtained from training on 262 and 33 attributes, respectively. Each multicolumn shows the performance achieved by one of the different composition operators presented in Section 4, as well as results obtained from predicting similarity on the basis of raw adjective (ADJ-only) and noun (N-only) vectors.

First and foremost, we observe best overall performance for the M&L model when combined with multiplicative vector composition ($\rho = 0.34$), even though the best results for this setting reported in M&L (2010) ($\rho = 0.46$) cannot be reproduced.

Nevertheless, the C-LDA models show a considerable performance improvement when the training material is constrained to appropriate attributes by an oracle (cf. Sect. 4). Another interesting observation is that the individual adjective and noun vectors produced by M&L and the C-LDA models, respectively, show diametrically opposed performance (cf. 3rd and 4th multicolumn in Table 1).

More in detail, C-LDA-A achieves relative improvements across all composition operators when comparing the 33-ATTR to the 262-ATTR setting. Contrasting C-LDA-A and C-LDA-T, the latter is clearly more effective on the larger training set, especially in combination with the × operator ($\rho = 0.28$). This might be due to the intersective character of multiplication, which requires densely populated components in both the adjective and the noun vector. This requirement meets best with the C-LDA-T model as long as the number of topics provided is large. The + operator, on the other hand, combines better with C-LDA-A. In the 33-ATTR setting, this combination even outperforms vector addition under the M&L model. Generally, C-LDA-A performs better on the smaller training set, where it leaves C-LDA-T behind in every configuration. This highlights that an interpretable, attribute-related meaning layer generalizing over latent topics can be effective if a small, discriminative set of attributes is available for training. Otherwise, C-LDA-T seems to be more powerful for the present similarity judgement task.

Analyzing the performance of the composition surrogates ADJ-only and N-only in the restricted 33-ATTR setting reveals an interesting twist in the quality of adjective vs. noun vectors: While M&L generally yields better results on noun vectors alone (as compared to adjective vectors), C-LDA-A clearly outperforms M&L in predicting similarity based on adjective meanings in isolation. In this configuration, M&L is also outperformed by the (very strong) dependency baseline which is, in turn, only slightly beaten by C-LDA-A in its best configuration. In fact, it is the ADJ-only surrogate under the C-LDA-A model in its best setting ($\rho = 0.29$) that comes closest to the overall best-performing M&L model. This indicates that modeling attributes in the latent semantics of adjectives can be informative for the present similarity prediction task. The poor quality of the noun vectors, however, limits the overall performance of the C-LDA models considerably.

Results on filtered test set. As can be seen from Table 2, our expectation that C-LDA-A and C-LDA-T should benefit from limiting the test set to instances related to attribute meanings is largely met. We observe overall improvement of correlation scores; also the characteristics of the individual models observed in Table 1 remain unchanged.

However, M&L benefits from filtering as well, and in some configurations, e.g. under vector addition, the relative improvement is even bigger for the latent word space models. This shows that M&L and our C-LDA models are not fully complementary, i.e. some aspects of attribute similarity are also covered by latent models.

Neverthelesss, the adjective/noun twist observed for individual vector performance is corroborated: C-LDA-A's adjective vectors outperform those of M&L by ten points (33 attributes, filtered setting; compared to six points on the complete test set), whereas the performance of the noun vectors drops even further. Again, the DepVSM baseline performs very strong on the adjective vectors in isolation, which clearly underlines that our dependency-based context selection procedure is effective. On the other hand, the individual noun vectors produced by M&L even yield the best overall result on the filtered test data, thus outperforming both composition methods.

Differences in adjective and noun vectors. In order to highlight qualitative differences of the individual adjective and noun vectors across the various models, we analyzed their informativeness in terms of entropy. The intuition is as follows: The lower the entropy exhibited by a vector, the more pronounced are its most prominent components. On the contrary, high entropy indicates a rather broad, less accentuated distribution of the probability mass over the vector components (cf. Hartung and Frank (2010)).

	262 attrs		33 attrs	
	avg	σ	avg	σ
C-LDA-A (JJ)	1.20	0.48	0.83	0.27
C-LDA-A (NN)	1.66	0.72	1.23	0.46
C-LDA-T (JJ)	0.92	0.04	0.50	0.04
C-LDA-T (NN)	1.10	0.06	0.60	0.02
M&L (JJ)	2.74	0.91	2.74	0.91
M&L (NN)	2.96	0.33	2.96	0.33
DepVSM (JJ)	0.48	0.61	0.65	0.32
DepVSM (NN)	0.38	0.67	0.96	0.21

Table 3: Average entropy of individual adjective and noun vectors across different models

The results of this analysis are displayed in Table 3. With regard to the C-LDA models, we observe lower entropy in adjective vectors compared to noun vectors across both training settings, which corresponds to their relative performance in the similarity prediction task. This indicates that C-LDA captures the relation between adjectives and attributes in a very pronounced way, and that this information proves valuable for similarity prediction.

The DepVSM model shows inconsistent results with regard to the different training sets. While the pattern observed for the C-LDA models is confirmed on the limited training set, training on the full set of 262 attributes results in more accentuated noun vectors. Given the huge standard deviations, however, we suppose that these figures are not very reliable.[5]

The correspondence between lower entropy and better performance we could observe for C-LDA-A and C-LDA-T is, however, not confirmed by the M&L word space model, as their adjective vectors exhibit lower entropy on average[6], while they persistently underperform relative to the noun vectors

[5] In fact, unlike the C-LDA models and M&L, DepVSM faces severe sparsity problems on the large training set, as becomes evident from the average total frequency mass per vector: Noun vectors accumulate 704 cooccurrence counts over 262 dimensions on average, while adjective vectors are populated with 1555 counts on average (652 vs. 1052 counts over 33 dimensions on the small training set).

[6] The entropy values of M&L are not directly comparable to those of the C-LDA models and DepVSM; M&L entropies are generally higher due to the higher dimensionality of the model.

(cf. Tables 1 and 2). Note, however, that the entropy values of individual adjective vectors disperse widely around the mean (σ=0.91). This suggests that a considerable proportion of M&L's adjective vectors is rather evenly distributed.

Analyzing the individual performance of noun vectors in terms of entropy is less conclusive. While the noun vectors consistently exhibit relatively high entropy, their varying performance across the different models cannot be explained. We hypothesize that the characteristics of the different models might be more decisive instead: Apparently, attributes as an abstract meaning layer are appropriate for modeling the contribution of adjectives to phrase similarity, whereas the contribution of nouns seems to be captured more effectively by M&L-like distributions along bags of context words.

6 Error Analysis

In order to gain deeper insight into the strengths and weaknesses of C-LDA-A and M&L, we extracted the ten most similar/dissimilar pairs (+Sim/$-$Sim$_{\text{C-LDA-A/M\&L}}$; cf. Table 4) according to system predictions, as well as the ten pairs on which system and human raters show highest/lowest agreement in terms of similarity scores (+Agr/$-$Agr$_{\text{C-LDA-A/M\&L}}$; cf. Table 5), for the best-performing model instance of C-LDA-A and M&L in the unfiltered 33-ATTR setting, respectively.

All pairs in +Sim$_{\text{C-LDA-A}}$ and +Sim$_{\text{M\&L}}$ exhibit matching attributes. +Sim$_{\text{C-LDA-A}}$ contains two pairs involving contrastive attribute values (vs. four in +Sim$_{\text{M\&L}}$): *long period – short time*, *hot weather – cold air*. Obviously, C-LDA-A is not prepared to recognize this type of dissimilarity, as it does not model the semantics and orientation of attribute *values*, and so assigns overly optimistic similarity rates. While this deficiency is explained for C-LDA, it is unexpected for M&L, where in +Sim$_{\text{M\&L}}$ we find pairs such as *old person – elderly lady* with similarity ratings that are almost identical to antonymous pairs discussed above, such as *high price – low cost*.

We further observe a striking difference regarding overall similarity ratings in both systems: We find high scores of 0.88 on average within +Sim$_{\text{C-LDA-A}}$, as opposed to 0.52 in +Sim$_{\text{M\&L}}$. The difference is less marked regarding $-$Sim. Similarly, we find overall low average similarity rates (0.2) in +Agr$_{\text{M\&L}}$, whereas +Agr$_{\text{C-LDA-A}}$ achieves somewhat higher rates (0.27). While all examples point towards dissimilarity, C-LDA-A shows more discriminative power, as exemplified by *hot weather – elderly lady* (lowest rating) vs. *central authority – local office* (highest rating). This suggests that, overall, C-LDA-A disposes of a more discriminative semantic representation to judge similarity – which of course can also go astray.

The disagreement set $-$Agr$_{\text{C-LDA-A}}$ contains the antonymous adjectives with high similarity ratings from +Sim$_{\text{C-LDA-A}}$, of course. We also note a high proportion (5/10) of pairs involving adjectives with vague and highly ambiguous attribute meanings, such as *good, new, certain, general*. These are difficult to capture, especially in combination with abstract noun concepts such as *information, effect* or *circumstance*.

An interesting type of similarity is represented by *early evening – previous day*. In this case, we observe a contrast in the semantics of the nouns involved, while the pair exhibits strong similarity on the attribute level, which is reflected in the system's similarity score. This type of similarity is reminiscent of relational analogies investigated in Turney (2008). A related example is *rural community – federal assembly*. Unlike the human judges, C-LDA predicts high similarity for both pairs.

The examples given in $-$Agr$_{\text{M\&L}}$, by contrast, clearly point to a lack in capturing adjective semantics, with misjudgements such as *effective way – efficient use*, *large number – vast amount* or *large quantity – great majority*.

Turning to $-$Agr$_{\text{C-LDA-A}}$ again, we find 9/10 items exhibit values greater than 0.67 (average: 0.78). This means the model yields a high number of false positives in rating similarity (with explanations and some reservations just discussed). All items in $-$Agr$_{\text{M\&L}}$, by contrast, have values below 0.36 (average: 0.16). That is, we again observe that this model assigns lower similarity scores. This is confirmed by a comparative analysis of average similarity scores on the entire test set: C-LDA-A;$+$ yields an average similarity of 0.48 (σ=0.05) over all instances, while M&L;$\times$ yields 0.16 on average (σ=0.16). The human ratings (after normalization to the scale from 0 to 1) amount to 0.39 (σ=0.26).

	SIMILARITY			
	C-LDA-A; +		M&L; ×	
+Sim	long period – short time	0.95	important part – significant role	0.66
	hot weather – cold air	0.95	certain circumstance – particular case	0.60
	different kind – various form	0.91	right hand – left arm	0.56
	better job – good place	0.89	long period – short time	0.55
	different part – various form	0.88	old person – elderly lady	0.54
	social event – special circumstance	0.88	high price – low cost	0.54
	better job – good effect	0.88	black hair – dark eye	0.48
	similar result – good effect	0.85	general principle – basic rule	0.44
	social activity – political action	0.82	special circumstance – particular case	0.43
	early evening – previous day	0.80	hot weather – cold air	0.43
−Sim	early stage – long period	0.11	old person – right hand	0.03
	northern region – early age	0.11	new information – further evidence	0.03
	earlier work – early evening	0.11	early stage – dark eye	0.01
	elderly woman – black hair	0.10	practical difficulty – cold air	0.01
	practical difficulty – cold air	0.08	left arm – elderly woman	0.01
	small house – old person	0.07	hot weather – elderly lady	0.00
	left arm – elderly woman	0.06	national government – cold air	0.00
	hot weather – further evidence	0.06	black hair – right hand	0.00
	dark eye – left arm	0.05	hot weather – further evidence	0.00
	national government – cold air	0.03	better job – economic problem	0.00

Table 4: Similarity scores predicted by optimal C-LDA-A and M&L model instances; 33-ATTR setting

	AGREEMENT			
	C-LDA-A; +		M&L; ×	
+Agr	major issue – american country	0.29	similar result – good effect	0.29
	efficient use – little room	0.29	small house – important part	0.14
	economic condition – american country	0.29	national government – new information	0.12
	public building – central authority	0.29	major issue – social event	0.26
	northern region – industrial area	0.28	new body – significant role	0.11
	new life – economic development	0.42	social event – special circumstance	0.25
	new body – significant role	0.13	economic development – rural community	0.32
	hot weather – elderly lady	0.13	new technology – public building	0.18
	social event – low cost	0.13	high price – short time	0.10
	central authority – local office	0.44	new body – whole system	0.24
−Agr	early evening – previous day	0.80	effective way – efficient use	0.29
	rural community – federal assembly	0.67	federal assembly – national government	0.24
	new information – general level	0.68	vast amount – high price	0.10
	similar result – good effect	0.85	different kind – various form	0.24
	better job – good effect	0.88	vast amount – large quantity	0.36
	social event – special circumstance	0.88	large number – vast amount	0.31
	better job – good place	0.89	older man – elderly woman	0.00
	certain circumstance – particular case	0.22	earlier work – early stage	0.00
	hot weather – cold air	0.95	large number – great majority	0.09
	long period – short time	0.95	large quantity – great majority	0.04

Table 5: Test pairs showing high and low agreement between systems and human raters, together with system similarity scores as obtained from optimal model instances; 33-ATTR setting

While these means are not fully comparable as they are the result of different composition operations, the standard deviations suggest that M&L's similarity predictions are dispersed over a larger range of the scale, while the C-LDA scores show only small variation. This missing spread might be one of the reasons for C-LDA's lower performance.

In summary, we note one obvious shortcoming in the C-LDA-A model, in that it does not capture dissimilarity due to distinct contrastive meanings of attribute values in cases of similarity on the noun and attribute levels. With its focus on attribute semantics, however, C-LDA-A is able to capture similarity due to *relational analogies*, as in *early evening – previous day* (0.8), whereas the latent model of M&L is clearly noun-oriented, and thus predicts a low similarity of 0.2 for this pair.

We conclude that the proposed attribute analysis of adjective-noun pairs implements an inherently relational form of similarity. Noun semantics is captured only indirectly, through the range of attributes found relevant for the noun. The current model also fully neglects the meaning of scalar attribute values. Whether a more comprehensive analysis of interpreted adjective-noun meanings is able to succeed in a paired similarity prediction task is an open issue to be explored in future work.

7 Conclusion

In this paper, we presented a distributional VSM that incorporates latent semantic information characterizing ontological attributes in the meaning of adjective-noun phrases, as obtained from C-LDA, a weakly supervised variant of LDA. Originally designed for an attribute selection task (Hartung and Frank, 2011), this model faces a true challenge when evaluated in a pairwise similarity judgement task against a high-dimensional word space model, such as M&L's VSM. In fact, our model is unable to compete with M&L even in its best configurations.

Thorough analysis reveals, however, that the quality of individual adjective and noun vectors is diametric across the two models: C-LDA, capitalizing on interpretable ontological dimensions, produces effective adjective vectors, whereas its noun representations lag behind. The inverse situation is observed for the word-based latent VSM of M&L.

One qualification is in order, though: In its current state, the C-LDA model relies on an "oracle" that pre-selects the attributes involved in the test set for the model to be trained on. Although one could argue that tailoring the context words to the target words has a similar effect in our re-implementation of M&L, interferences of this kind are not desirable in principle. Future work will need to explore in more detail possible attribute ranges with regard to their usefulness for different tasks and data sets.

Our comparative investigaton of the specific strengths and weaknesses of the models indicates that they focus on different aspects of similarity: M&L, possibly due to its higher and more discriminative dimensionality, tends to produce more efficient noun vectors. Overall, this model accords better with human similarity judgements across diverse aspects of similarity than the more focused attribute-oriented LDA models. The C-LDA models focus on a specific, interpretable meaning dimension shared by adjectives and nouns, with a tendency for stronger modeling capacity for adjectives. They are currently not prepared to capture dissimilarity in cases of contrastive attribute values, while on the positive side, they effectively cope with relational analogies, both with similar and dissimilar noun meanings.

Our findings suggest that adding more discriminative power to the noun representations and scalar information about attribute values to the adjective vectors might be beneficial. Further research is needed to investigate how to combine interpretable semantic representations tailored to specific relations, as captured by C-LDA, with M&L-like bag-of-words representations in a single distributional model.

Applying interpreted models to the present similarity rating task will still remain a challenge, as it involves mapping diverse mixtures of aspects and grades of similarity to human judgements. However, if the performance of an integrated model can compete with a purely latent semantic analysis, this offers a clear advantage for more general tasks that require linking phrase meaning to symbolic knowledge bases such as (multilingual) ontologies, or for application scenarios that involve discrete semantic labels, such as text classification based on topic modeling (Blei et al., 2003) or fine-grained named entity classification (Ekbal et al., 2010).

References

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

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.

Marco Baroni, 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.

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

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

Asif Ekbal, Eva Sourjikova, Anette Frank, and Simone Ponzetto. 2010. Assessing the Challenge of Fine-grained Named Entity Recognition and Classification. In *Proceedings of the ACL 2010 Named Entity Workshop (NEWS)*, Uppsala, Sweden.

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.

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. Exploring Supervised LDA Models for Assigning Attributes to Adjective-Noun Phrases. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, UK.

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.

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.

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.

Peter D. Turney. 2008. A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 905–912, Manchester, UK.