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

Matthias Hartung **Anette Frank**
Computational Linguistics Department
Heidelberg University

GEMS 2011
Edinburgh, July 31

Motivation: "Use Cases" of Distributional Models

Distributional Similarity

- distributional models provide graded *similarity* judgements for word or phrase pairs
- *sources* of similarity are usually disregarded
- desirable goal: predict degree of similarity **and** its source

Example:

elderly lady vs. *old woman*

- high degree of similarity
- primary source of similarity: shared feature AGE

Distributional Models in Categorial Prediction Tasks

Example: Attribute Selection

- What are the **attributes** of a concept that are highlighted in an adjective-noun phrase ?
- well-known problem in formal semantics:
  - *short hair* → LENGTH
  - *short discussion* → DURATION
  - *short flight* → DISTANCE or DURATION
- Hartung & Frank (2010): formulate attribute selection as a **compositional process** in distributional framework

Attribute Selection: Previous Work

Pattern-based VSM: Hartung & Frank (2010)

	COLOR	DIRECT.	DURAT.	SHAPE	SIZE	SMELL	SPEED	TASTE	TEMP.	WEIGHT
enormous	1	1	0	1	45	0	4	0	0	21
ball	14	38	2	20	26	0	45	0	0	20
enormous $\times$ *ball*	14	38	0	20	**1170**	0	180	0	0	420
enormous + *ball*	15	39	2	21	**71**	0	49	0	0	41

- vector component values: raw corpus frequencies obtained from lexico-syntactic patterns such as
  - (A1) `ATTR of DT? NN is|was JJ`
  - (N2) `DT ATTR of DT? RB? JJ? NN`
- restriction to 10 manually selected attribute nouns
- sparsity of patterns; to be alleviated by integration of LDA topic models

Focus of Today's Talk

Is a distributional model tailored to *attribute selection* **effective in** *similarity prediction* **?**

Approach:

- construct *attribute-related meaning representations* (AMRs) for adjectives and nouns in a distributional model (incorporating LDA topic models)
- comparison against *latent* VSM of Mitchell & Lapata (2010; henceforth: **M&L**) on similarity judgement data

Outline

Using LDA for Lexical Semantics

LDA in Document Modeling

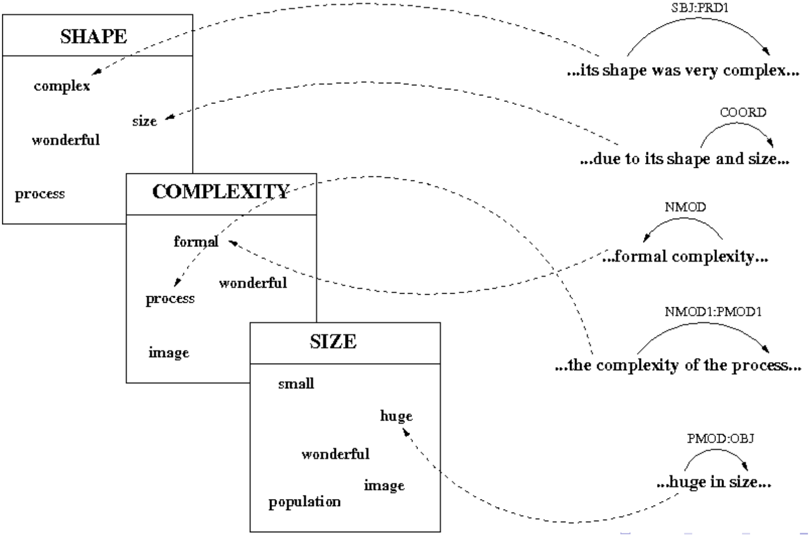
- hidden variable model for *document modeling*
- decompose document collection into *topics* that capture their *latent semantics* in a more abstract way than BOWs

Porting LDA to Attribute Semantics

- build "pseudo-documents" as distributional profiles of attribute meaning
- resulting topics are highly "attribute-specific"
- similar approaches in other areas of lexical semantics:
  - semantic relation learning (Ritter et al., 2010)
  - selectional preference modeling (Ó Séaghdha, 2010)
  - word sense disambiguation (Li et al., 2010)

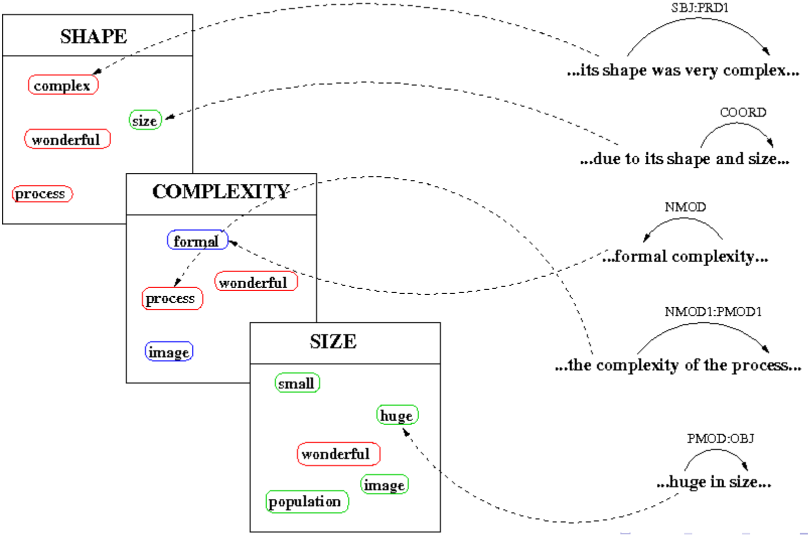
Attribute Modeling by Controled LDA (C-LDA)

Constructing "Pseudo-Documents":



Attribute Modeling by Controled LDA (C-LDA)

Constructing "Pseudo-Documents":



C-LDA: Generative Process

For each topic $k \in \{1, \ldots, K\}$:
    Generate $\beta_k \sim Dir_V(\eta)$
For each document d:
    Generate $\theta_d \sim Dir(\alpha)$
    For each n in $\{1, \ldots, N_d\}$:
        Generate $z_{d,n} \sim Mult(\theta_d)$ with $z_{d,n} \in \{1, \ldots, K\}$
        Generate $w_{d,n} \sim Mult(\beta_{z_{d,n}})$ with $w_{d,n} \in \{1, \ldots, V\}$

(Blei et al., 2003)

Integrating Attribute Models into the VSM Framework (I)

C-LDA-A: Attributes as Meaning Dimensions

	COLOR	DIRECT.	DURAT.	SHAPE	SIZE	SMELL	SPEED	TASTE	TEMP.	WEIGHT
hot	18	3	1	4	1	14	1	5	174	3
meal	3	5	119	10	11	5	4	103	3	33
hot $\times$ meal	0.05	0.02	0.12	0.04	0.01	0.07	0.00	0.51	0.52	0.10
hot $+$ meal	21	8	120	14	11	19	5	108	177	36

Table: VSM with C-LDA probabilities (scaled by 10^3)

Setting Vector Component Values:

$$v_{\langle w,a \rangle} = P(w|a) \approx P(w|d_a) = \sum_t P(w|t)P(t|d_a)$$

Integrating Attribute Models into the VSM Framework (II)

C-LDA-T: Topics as Meaning Dimensions

	TOPIC 1	TOPIC 2	TOPIC 3	TOPIC 4	TOPIC 5	TOPIC 6	TOPIC 7	TOPIC 8	TOPIC 9	TOPIC 10
hot	27	4	1	14	3	14	0	9	34	3
meal	62	10	82	11	12	8	4	14	77	33
hot $\times$ meal	1.67	0.04	0.08	0.15	0.04	0.11	0.00	0.13	2.62	0.10
hot + meal	89	14	83	25	15	22	4	23	111	36

Table: VSM with C-LDA probabilities (scaled by 10^3)

Setting Vector Component Values:

$$v_{\langle w,t \rangle} = P(w|t)$$

Integrating Attribute Models into the VSM Framework (III)

Vector Composition Operators:

- vector multiplication ($\times$)
- vector addition ($+$)

(Mitchell & Lapata, 2010)

"Composition Surrogates":

- ADJ-only: take adjective vector instead of composition
- N-only: take noun vector instead of composition

(Hartung & Frank, 2010)

Taking Stock...

Models for Similarity Prediction

Attribute-specific Models:

- **C-LDA-A:** *attributes* as interpreted dimensions
- **C-LDA-T:** attribute-related *topics* as dimensions

Latent Model:

- **M&L:** 5w+5w context windows, 2000 most frequent *context words* as dimensions (Mitchell & Lapata, 2010)

Experimental Settings (I)

Training Data for C-LDA Models:

- **Complete Attribute Set:** 262 attribute nouns linked to at least one adjective by the `attribute` relation in WordNet
- **"Attribute Oracle":** 33 attribute nouns linked to one of the adjectives occurring in the M&L test set

Testing Data:

- **Complete Test Set:** all 108 pairs of adj-noun phrases contained in the M&L benchmark data
- **Filtered Test Set:** 43 pairs of adj-noun phrases from M&L where both adjectives bear an attribute meaning according to WordNet

Experimental Settings (II)

Evaluation Procedure:

1. compute *cosine similarity* between the composed vectors representing the adjective-noun phrases in each test pair
2. measure *correlation* between model scores and human judgements in terms of Spearman's ρ; treat each human rating as an individual data point

Experimental Results (I)

Complete Test Set:

		+		$\times$		ADJ-only		N-only	
		avg	best	avg	best	avg	best	avg	best
262 attrs	C-LDA-A	0.19	0.25	0.15	0.20	0.17	0.23	0.11	0.23
	C-LDA-T	0.19	0.24	0.28	0.31	0.20	0.24	0.18	0.24
	M&L	0.21		**0.34**		0.19		**0.27**	
33 attrs	C-LDA-A	**0.23**	0.27	0.21	0.24	**0.27**	0.29	0.17	0.22
	C-LDA-T	0.21	0.28	0.14	0.23	0.22	0.27	0.10	0.21
	M&L	0.21		**0.34**		0.19		**0.27**	

- M&L$_\times$ performs best in both training scenarios
- C-LDA models generally benefit from confined training data (except for C-LDA-T$_\times$)
- individual adjective and noun vectors produced by M&L and the C-LDA models show diametrically opposed performance

Experimental Results (II)

Filtered Test Set (Attribute-related Pairs only):

		+		×		ADJ-only		N-only	
		avg	best	avg	best	avg	best	avg	best
262 attrs	C-LDA-A	0.22	0.31	0.12	0.30	0.18	0.30	0.17	0.28
	C-LDA-T	0.25	0.30	0.26	0.35	0.24	0.29	0.19	0.23
	M&L	**0.38**		**0.40**		0.24		**0.43**	
33 attrs	C-LDA-A	0.29	0.32	0.31	0.36	0.34	0.38	0.09	0.18
	C-LDA-T	0.26	0.36	0.14	0.30	0.28	0.38	0.03	0.18
	M&L	**0.38**		**0.40**		0.24		**0.43**	

- improvements of C-LDA models on restricted test set: C-LDA is informative for attribute-related test instances
- relative improvements of M&L are even higher than those of C-LDA in some configurations
- adjective/noun twist is corroborated

Differences between Adjective and Noun Vectors

	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	

Table: Avg. entropy of adj. and noun vectors

- **hypothesis:** *information* in adjective and noun vectors mirrors their relative performance
- low entropy $\equiv$ high information, and vice versa

- hypothesis confirmed for C-LDA only
- M&L: diametric pattern, but considerable proportion of relatively uninformative adjective vectors (cf. σ=0.91)

Qualitative Analysis (I)

System Predictions: Most Similar/Dissimilar Pairs

	C-LDA-A; $+$		M&L; $\times$	
$+$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
$-$Sim	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: Similarity scores predicted by C-LDA-A (optimal) and M&L; 33 attrs

- large majority of pairs in $+\text{Sim}_{\text{C-LDA-A}}$ and $+\text{Sim}_{\text{M\&L}}$ represent matching attributes
- both models cannot deal with *antonymous* attribute *values*
- C-LDA-A utilizes larger range on the similarity scale

Qualitative Analysis (I)

System Predictions: Most Similar/Dissimilar Pairs

	C-LDA-A; $+$		M&L; $\times$	
$+$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
$-$Sim	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: Similarity scores predicted by C-LDA-A (optimal) and M&L; 33 attrs

- large majority of pairs in $+\text{Sim}_{\text{C-LDA-A}}$ and $+\text{Sim}_{\text{M\&L}}$ represent matching attributes
- both models cannot deal with *antonymous* attribute *values*
- C-LDA-A utilizes larger range on the similarity scale

Qualitative Analysis (I)

System Predictions: Most Similar/Dissimilar Pairs

	C-LDA-A; $+$		M&L; $\times$	
+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
−Sim	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: Similarity scores predicted by C-LDA-A (optimal) and M&L; 33 attrs

- large majority of pairs in $+\text{Sim}_{\text{C-LDA-A}}$ and $+\text{Sim}_{\text{M\&L}}$ represent matching attributes
- both models cannot deal with *antonymous* attribute *values*
- C-LDA-A utilizes larger range on the similarity scale

Qualitative Analysis (I)

System Predictions: Most Similar/Dissimilar Pairs

	C-LDA-A; $+$		**M&L;** $\times$	
$+$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
$-$Sim	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: Similarity scores predicted by C-LDA-A (optimal) and M&L; 33 attrs

- large majority of pairs in $+\text{Sim}_{\text{C-LDA-A}}$ and $+\text{Sim}_{\text{M\&L}}$ represent matching attributes
- both models cannot deal with *antonymous* attribute *values*
- C-LDA-A utilizes larger range on the similarity scale

Qualitative Analysis (II)

Agreement between Systems and Human Judgements

	C-LDA-A; $+$		M&L; $\times$	
+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
−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

Table: High and low agreement pairs (systems vs. human raters), together with system similarity scores as obtained from optimal model instances; 33 attrs

- $-\text{Agr}_{\text{C-LDA-A}}$: many adjectives with general or vague attribute meanings in combination with abstract nouns
- $-\text{Agr}_{\text{M\&L}}$: lack of attribute-related adjective semantics
- notion of similarity underlying C-LDA-A is close to *relational analogies* (Turney, 2008)

Qualitative Analysis (II)

Agreement between Systems and Human Judgements

	C-LDA-A; $+$		**M&L**; $\times$	
$+$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
$-$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

Table: High and low agreement pairs (systems vs. human raters), together with system similarity scores as obtained from optimal model instances; 33 attrs

- –$\text{Agr}_{\text{C-LDA-A}}$: many adjectives with general or vague attribute meanings in combination with abstract nouns
- –$\text{Agr}_{\text{M\&L}}$: lack of attribute-related adjective semantics
- notion of similarity underlying C-LDA-A is close to *relational analogies* (Turney, 2008)

Qualitative Analysis (II)

Agreement between Systems and Human Judgements

	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
−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

Table: High and low agreement pairs (systems vs. human raters), together with system similarity scores as obtained from optimal model instances; 33 attrs

- –Agr$_{\text{C-LDA-A}}$: many adjectives with general or vague attribute meanings in combination with abstract nouns
- –Agr$_{\text{M\&L}}$: lack of attribute-related adjective semantics
- notion of similarity underlying C-LDA-A is close to *relational analogies* (Turney, 2008)

Qualitative Analysis (II)

Agreement between Systems and Human Judgements

	C-LDA-A; $+$		M&L; $\times$	
$+$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
$-$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

Table: High and low agreement pairs (systems vs. human raters), together with system similarity scores as obtained from optimal model instances; 33 attrs

- –$\text{Agr}_{\text{C-LDA-A}}$: many adjectives with general or vague attribute meanings in combination with abstract nouns
- –$\text{Agr}_{\text{M\&L}}$: lack of attribute-related adjective semantics
- notion of similarity underlying C-LDA-A is close to *relational analogies* (Turney, 2008)

Conclusions and Outlook (I)

Contributions:

- approach to integrate attribute-specific topic models into distributional VSM framework
- assessed feasibility of similarity prediction along interpretable dimensions of meaning

Findings:

1. **C-LDA-A vs. C-LDA-T:**
   - C-LDA-T performs better on the full data set
   - C-LDA-A is advantageous on attribute-related subset
2. **C-LDA vs. M&L:**
   - lower overall performance of C-LDA models
   - models capture different types of similarity
   - diametric strengths and weaknesses: individual adjective vectors of C-LDA outperform those of M&L; nouns lag behind

Conclusions and Outlook (II)

Future Work:

- more thorough analysis of different shades of similarity underlying the data
- enrich noun representations of C-LDA models
- integrate semantics for attribute values
- possibly combine latent and interpretable models ?

References

- Baroni, Marco, Silvia Bernardini, Adriano Ferraresi and Eros Zanchetta (2009): The WaCky Wide Web. A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation 43(3)*: 209-226.
- Blei, David, Andrew Ng and Michael Jordan (2003): Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*: 993-1022.
- Hartung, Matthias & Anette Frank (2010): A Structured Vector Space Model for Hidden Attribute Meaning in Adjective-Noun Phrases. Proceedings of COLING, Beijing, China: 430-438.
- Li, Linlin, Benjamin Roth & Caroline Sporleder (2010): Topic models for word sense disambiguation and token-based idiom detection. Proceedings of ACL: 1138-1147.
- Mitchell, Jeff & Mirella Lapata (2009): Language Models Based on Semantic Composition. Proceedings of EMNLP, Singapore: 430-439.
- Mitchell, Jeff & Mirella Lapata (2010): Composition in Distributional Models of Semantics. *Cognitive Science 34(8)*: 1388-1429.
- Ó Séaghdha, Diarmuid (2010): Latent Variable Models of Selectional Preference. Proceedings of ACL: 435-444.
- Ritter, Alan, Mausam & Oren Etzioni (2010): A Latent Dirichlet Allocation Method for Selectional Preferences. Proceedings of ACL: 424-434.

Thanks...

...for your attention.
Questions ?