Potential and Limits of Distributional Approaches to Semantic Relatedness

PD Dr. Sabine Schulte im Walde

Institut für Maschinelle Sprachverarbeitung (IMS)
Universität Stuttgart
Germany

Joint Symposium on Semantic Processing (JSSP)
FBK, Trento
November 21, 2013

Semantics in Corpus Distributions

BLA sledge BLA BLA BLA BLA
BLA BLA BLA snow BLA BLA
BLA BLA white BLA BLA BLA
BLA BLA BLA BLA BLA winter

Research Questions

1. Distributional Information
    - potential and limits
    - extensions and alternatives

Research Questions

1. Distributional Information
   - potential and limits
   - extensions and alternatives

2. Salient Distributional Features
   - default features
   - phenomenon-related features

Research Questions

1. Distributional Information
   - potential and limits
   - extensions and alternatives
2. Salient Distributional Features
   - default features
   - phenomenon-related features
3. Ambiguity in Vector Spaces
   - vector spaces summarise over senses
   - definition of vector regions
   - characterisation of (regular) polysemy
   - identification of polysemous objects

Phenomena

- Semantic Relatedness
  agreement on semantic properties of words and phrases

Phenomena

- Semantic Relatedness
  agreement on semantic properties of words and phrases

- Phenomena:

  1. paradigmatic semantic relations (German, English, Italian)
  2. compositionality of German noun-noun compounds
  3. senses and polysemy of German prepositions

Phenomena

- Semantic Relatedness
  agreement on semantic properties of words and phrases

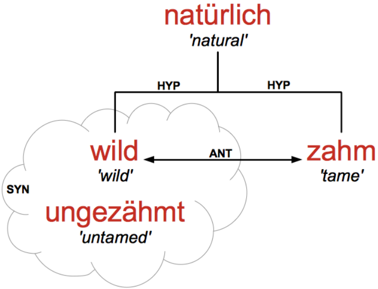
- Phenomena:

  1. paradigmatic semantic relations (German, English, Italian)
  2. compositionality of German noun-noun compounds
  3. senses and polysemy of German prepositions

- Research Methodology:

  - interdisciplinary framework: linguistics, cognition, computation
  - distributional information at the syntax-semantics interface
  - unsupervised machine learning approaches
  - extrinsic evaluation: statistical machine translation

Paradigmatic Semantic Relations



Dataset

- Task: distinguish between paradigmatic semantic relation pairs
  'The boy/girl/person loves/hates the cat.'

Dataset

- Task: distinguish between paradigmatic semantic relation pairs
  'The boy/girl/person loves/hates the cat.'
- Languages: German, English, Italian (Stuttgart; Pisa)
- Relations: synonymy, antonymy, hypernymy, co-hyponymy
- Word Classes: nouns, verbs, adjectives
- Dataset: random choice of 99 WordNet targets per word class
  - frequency class (low; mid; high)
  - polysemy class (monosemous; two senses; $>$2 senses)
  - size of semantic class
- Experiments: generation and rating of pairs, using AMT
  (Scheible & Schulte im Walde, in preparation)

German Examples

Generation:

	ANT		SYN		HYP	
NOUN	*Bein/Arm* (leg/arm)	10	*Killer/Mörder* (killer)	8	*Ekel/Gefühl* (disgust/feeling)	7
	Zeit/Raum (time/space)	3	*Gerät/Apparat* (device)	3	*Arzt/Beruf* (doctor/profession)	5
VERB	*verbieten/erlauben* (forbid/allow)	10	*üben/trainieren* (practise)	6	*trampeln/gehen* (lumber/walk)	6
	setzen/stehen (sit/stand)	4	*setzen/platzieren* (place)	3	*wehen/bewegen* (wave/move)	3
ADJ	*dunkel/hell* (dark/light)	10	*mild/sanft* (smooth)	9	*grün/farbig* (green/colourful)	5
	heiter/trist (cheerful/sad)	2	*bekannt/vertraut* (familiar)	4	*heiter/hell* (bright/light)	1

German Examples

Generation:

	ANT		SYN		HYP	
NOUN	*Bein/Arm* (leg/arm)	10	*Killer/Mörder* (killer)	8	*Ekel/Gefühl* (disgust/feeling)	7
	Zeit/Raum (time/space)	3	*Gerät/Apparat* (device)	3	*Arzt/Beruf* (doctor/profession)	5
VERB	*verbieten/erlauben* (forbid/allow)	10	*üben/trainieren* (practise)	6	*trampeln/gehen* (lumber/walk)	6
	setzen/stehen (sit/stand)	4	*setzen/platzieren* (place)	3	*wehen/bewegen* (wave/move)	3
ADJ	*dunkel/hell* (dark/light)	10	*mild/sanft* (smooth)	9	*grün/farbig* (green/colourful)	5
	heiter/trist (cheerful/sad)	2	*bekannt/vertraut* (familiar)	4	*heiter/hell* (bright/light)	1

Rating:

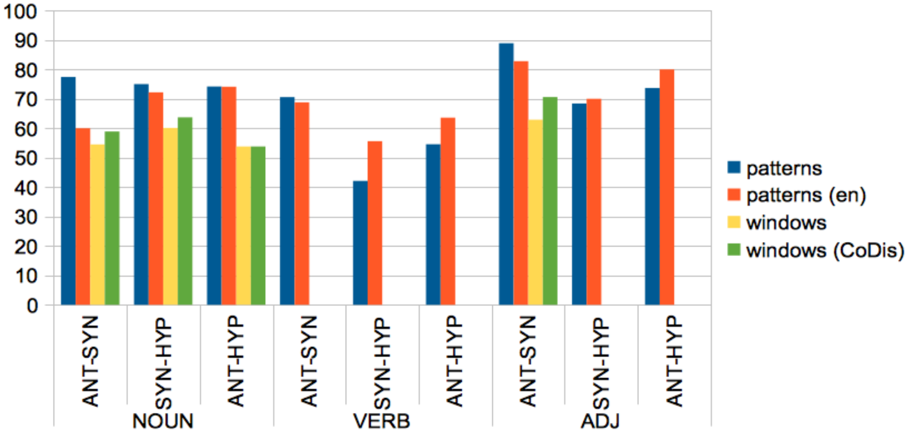
	Target	Generation	ANT	SYN	HYP
NOUN	*Zeit/Raum* (time/space)	ANT: 3	4.6	1.4	1.5
	Gerät/Maschine (device/machine)	SYN: 5 HYP: 2	1.0	4.7	3.4

Distributional Models

- Pattern-based Features
  (Schulte im Walde & Köper, 2013; Nayak, Internship 2012)
  - standard lexico-syntactic patterns
  - variations: frequency; length; specificity; reliability
  - nearest-centroid classification

- Window Co-Occurrence Features
  (Müller, Scheible, Schulte im Walde; IJCNLP, 2013)
  - standard similarity in co-occurrence
  - window sizes 5 and 20 (left and right)
  - contribution of parts-of-speech of co-occurring words
  - simple context disambiguation (CoDis)

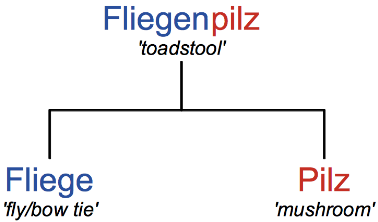
Results


100
90
80
70
60
50
40
30
20
10
0
ANT-SYN
SYN-HYP
ANT-HYP
NOUN
VERB
ADJ
patterns
patterns (en)
windows
windows (CoDis)

Insights

1. Distributional Information
   - standard approaches outperform baselines significantly
   - success varies wrt word classes and relations
2. Salient Distributional Features
   - patterns outperform windows
   - large-scale, noisy patterns perform best
   - different effect of co-occurring word classes wrt target word classes and relation types: V for ADJ; ADJ/V for N
3. Ambiguity in Vector Spaces
   - CoDis features disambiguate relation pair senses

German Noun-Noun Compounds



Dataset

- Composition:
  - 244 concrete, depictable German noun-noun compounds; subset of von der Heide & Borgwaldt (2009)
  - compounds, modifiers and heads are nouns
  - four compositionality classes (O=opaque; T=transparent): O+O, T+T, O+T, T+O
- Examples:
  - *Postbote* 'post man': *Post* 'mail' + *Bote* 'messenger'
  - *Löwenzahn* 'dandelion': *Löwe* 'lion' + *Zahn* 'tooth'
  - *Fliegenpilz* 'toadstool': *Fliege* 'fly/bow tie' + *Pilz* 'mushroom'
  - *Feuerzeug* 'lighter': *Feuer* 'fire' + *Zeug* 'stuff'

Examples

Human ratings on the degree of compositionality:

- compound 'whole' ratings
- compound–constituent ratings

Examples

Human ratings on the degree of compositionality:

- compound 'whole' ratings
- compound–constituent ratings

Compounds			Mean Ratings and Standard Deviations		
whole	literal meanings of constituents		whole	modifier	head
Ahornblatt 'maple leaf'	maple	leaf	6.03 ± 1.49	**5.64** ± 1.63	**5.71** ± 1.70
Postbote 'post man'	mail	messenger	6.33 ± 0.96	**5.87** ± 1.55	**5.10** ± 1.99
Seezunge 'sole'	sea	tongue	1.85 ± 1.28	**3.57** ± 2.42	**3.27** ± 2.32
Windlicht 'storm lamp'	wind	light	3.52 ± 2.08	**3.07** ± 2.12	**4.27** ± 2.36
Löwenzahn 'dandelion'	lion	tooth	1.66 ± 1.54	**2.10** ± 1.84	**2.23** ± 1.92
Maulwurf 'mole'	mouth	throw	1.58 ± 1.43	**2.21** ± 1.68	**2.76** ± 2.10
Fliegenpilz 'toadstool'	fly/bow tie	mushroom	2.00 ± 1.20	**1.93** ± 1.28	**6.55** ± 0.63
Flohmarkt 'flea market'	flea	market	2.31 ± 1.65	**1.50** ± 1.22	**6.03** ± 1.50
Feuerzeug 'lighter'	fire	stuff	4.58 ± 1.75	**5.87** ± 1.01	**1.90** ± 1.03
Fleischwolf 'meat chopper'	meat	wolf	1.70 ± 1.05	**6.00** ± 1.44	**1.90** ± 1.42

Models

1. Distributional model of lexical, corpus-based co-occurrence (Schulte im Walde et al., 2013):
   - Task: predict the degree of compositionality of the compounds
   - Subtask 1: compare window-based vs. syntax-based features
   - Subtask 2: compare contributions of modifiers vs. heads

2. Multi-modal LDA model incorporating lexical data (co-occurrence), experiential data (associations, features) and visual data (pictures); Roller & Schulte im Walde (2013)
   - Task: predict the degree of compositionality of the compounds

Results

- Nouns provide most salient features: $\rho = .6497$ (window: 20)
- Window-based features outperform syntax-based features

- Salient features to predict similarities between compound–modifier vs. compound–head pairs are different: small windows: compound–head $>$ compound–modifier; syntactic features: compound–head $>$ compound–modifier

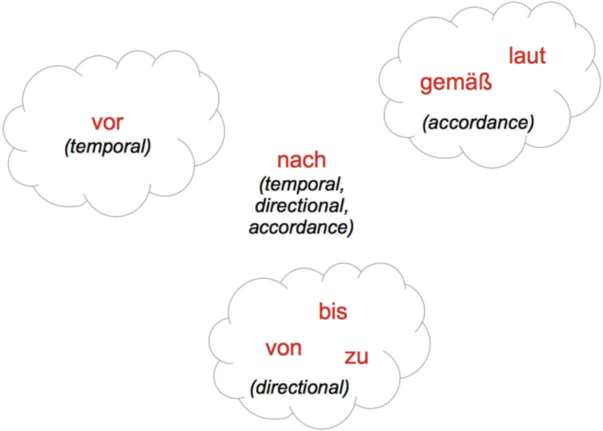
- Influence of modifier meaning on compound meaning is stronger than influence of head meaning

- Hybrid LDA model concatenating textual features, association norms, SURF features and GIST clusters outperforms textual model and various 2- and 3-dimensional LDA models

Insights

1. Distributional Information
   - window information outperforms syntactic information
   - distributional model outperforms multi-modal model
2. Salient Distributional Features
   - nouns in 20-word windows
   - differ wrt compound–modifer vs. compound–head predictions
3. Ambiguity in Vector Spaces
   - not yet resolved

Polysemy of German Prepositions



Dataset

- German prepositions are notoriously ambiguous:
  nach drei Stunden/Berlin/Meinung
  'after three hours/to Berlin/according to'

- Tasks:
  1. cluster prepositions into senses
  2. identify polysemous prepositions

- Sources for preposition senses:
  grammar books; gold standards from earlier projects

Framework

Feature-based setting (Springorum, Schulte im Walde, Utt, 2013)

1. Associate prepositions with a distributional feature set.
2. Perform hard clustering using Self-Organising Maps.
3. Transfer hard clusterings to soft clusterings.
4. Explore and evaluate cluster analyses.

Rank-based setting (Köper & Schulte im Walde, submitted)

1. Associate prepositions with a distributional feature set.
2. Calculate similarity ranks of preposition pairs.
3. Sort or cluster prepositions into monosemous vs. polysemous.

Hypotheses

What are the spatial properties of polysemous objects?

Hypotheses

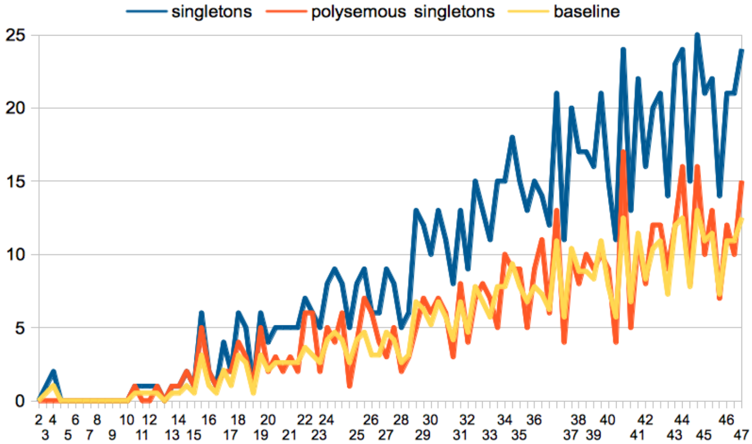
What are the spatial properties of polysemous objects?

Alternative hypotheses, so far:

- Singletons represent polysemy.
- Polysemous prepositions are misclassified.
- Cluster membership rate corresponds to ambiguity rate.
- Polysemous prepositions are similar to many prepositions.

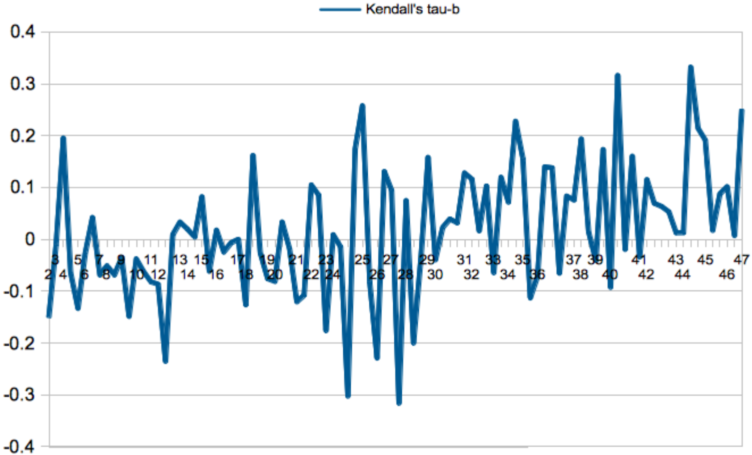
Singletons represent Polysemy

Number of singletons (containing polysemous prepositions):



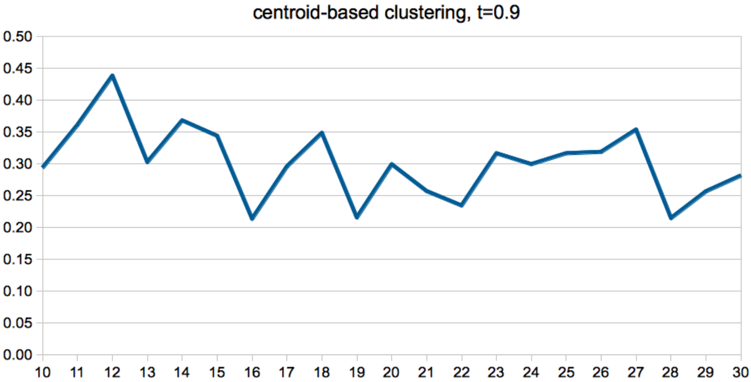
Polysemous Prepositions are Misclassified

Correlation of Silhouette Value and preposition ambiguity rate:

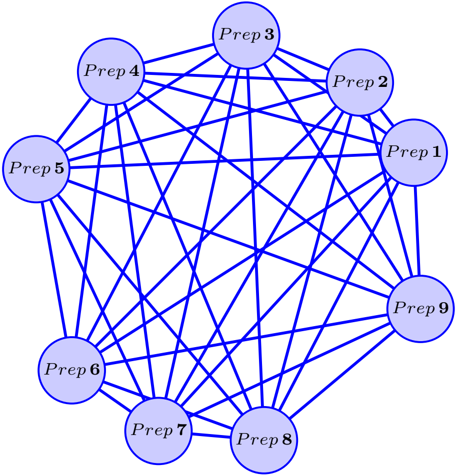


Polysemous Prepositions and Cluster Assignment

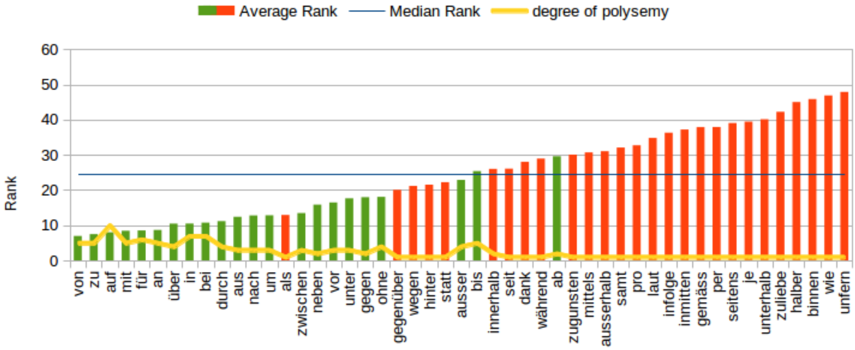
Correlation of cluster membership rate and ambiguity rate:



Similarity-based Rank Values



Similarity-Rank-based Identification of Polysemy



Insights

1. Distributional Information
    - standard dependency features allow a reasonable classification
    - distributional information distinguishes monosemous and polysemous prepositions
2. Salient Distributional Features
    - subcategorised nouns distinguish preposition senses
    - a similarity-based ranking relying on binary features distinguishes monosemous from polysemous prepositions
3. Ambiguity in Vector Spaces
    - first step towards identifying ambiguous objects

Distributional Information for SMT

- Hierarchical machine translation system
- Two-step translation procedure: (i) build translation system on stemmed representations; (ii) inflect translation
- Example for case confusion in English–German SMT:

input		[why]$_1$ [the government]$_2$ [ordered]$_3$ [the ongoing military actions]$_4$
output	stemmed	[warum]$_1$ [d Regierung]$_2$ [d anhaltend militärisch Aktion]$_4$ [angeordnet]$_3$
	inflected	[warum]$_1$ [**die** Regierung]$_2$ [**der** anhalten**den** militärisch**en** Aktion**en**]$_4$ [angeordnet]$_3$

- Integration of subcategorisation information:
  - features on source-side syntactic subcategorisation
  - external knowledge base with quantitative, dependency-based information about target-side subcategorisation frames
- Evaluation shows positive impact on translation quality

Summary and Conclusions

1. Distributional Information
    - distinguishes between paradigmatic relations
    - predicts the compositionality of noun-noun compounds
    - (identifies polysemous prepositions and preposition senses)
    - is useful for statistical machine translation

Summary and Conclusions

1. Distributional Information
   - distinguishes between paradigmatic relations
   - predicts the compositionality of noun-noun compounds
   - (identifies polysemous prepositions and preposition senses)
   - is useful for statistical machine translation
2. Salient Distributional Features
   - default features might represent a first step but ...
   - phenomenon-related features tell the linguistic story

Summary and Conclusions

1. Distributional Information
   - distinguishes between paradigmatic relations
   - predicts the compositionality of noun-noun compounds
   - (identifies polysemous prepositions and preposition senses)
   - is useful for statistical machine translation
2. Salient Distributional Features
   - default features might represent a first step but ...
   - phenomenon-related features tell the linguistic story
3. Ambiguity in Vector Spaces
   - CoDis is a simple but effective approach to disambiguate pair-based ambiguity
   - spatial location of polysemous objects: needs more exploration