Towards a Lexicologically Informed Parameter Evaluation of Distributional Modelling in Lexical Semantics

Thomas Wielfaert, Kris Heylen, Jocelyne Daems,
Dirk Speelman & Dirk Geeraerts

KU Leuven
Quantitative Lexicology and Variational Linguistics

Purpose of the talk

THEORETICAL

- Study the structure of lexical variation: mapping of meaning onto lexemes in different varieties.
- Analyse how this structure is apparent in usage data

METHODOLOGICAL

- Semantic Vector Spaces as a method for the quantitative, large-scale, corpus-based analysis of lexical semantics
- Interactive Visualisation of distributional models as an exploratory, visual analytic tool for lexicology
- Creating a 'gold standard' and cluster evaluation.

Overview

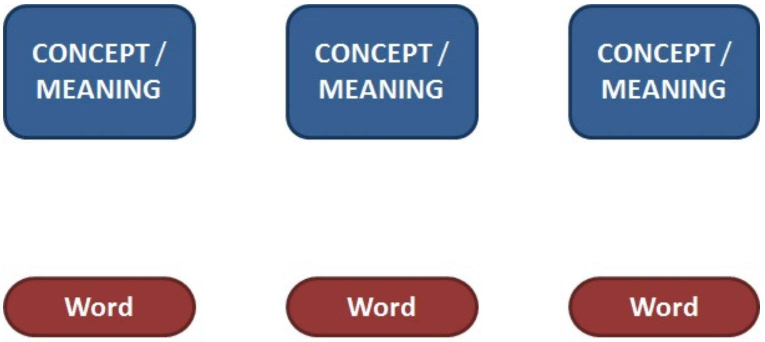
1. Linguistic Background
2. Semantic Vector Spaces
3. Visual Analytics
4. Creating a 'gold standard' and cluster evalution.
5. Discussion and future work

Overview

1. Linguistic Background
2. Semantic Vector Spaces
3. Visual Analytics
4. Creating a 'gold standard' and cluster evalution.
5. Discussion and future work

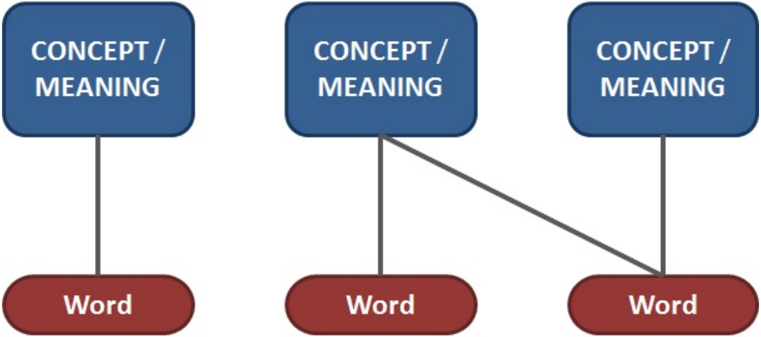
Linguistic Background

Structure of Lexical Variation (Geeraerts et al. 1994):



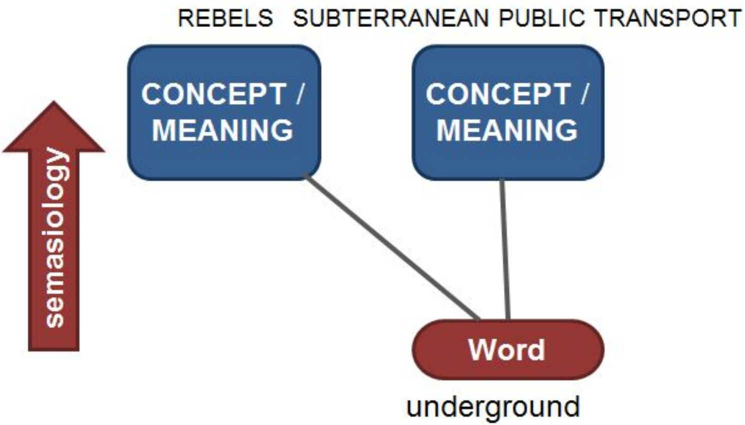
Linguistic Background

Structure of Lexical Variation (Geeraerts et al. 1994):



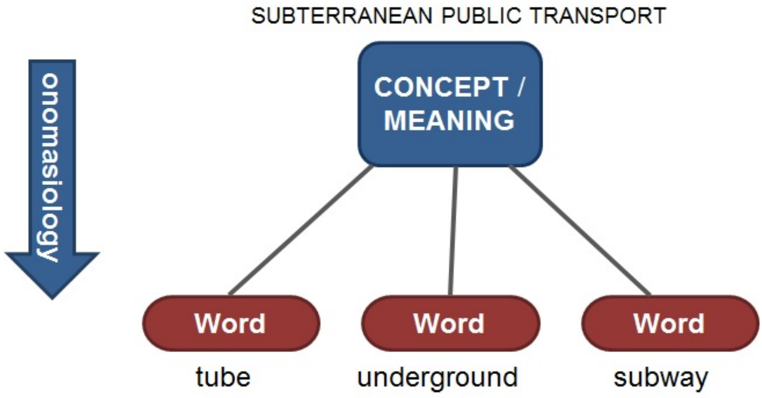
Linguistic Background

Structure of Lexical Variation (Geeraerts et al. 1994):



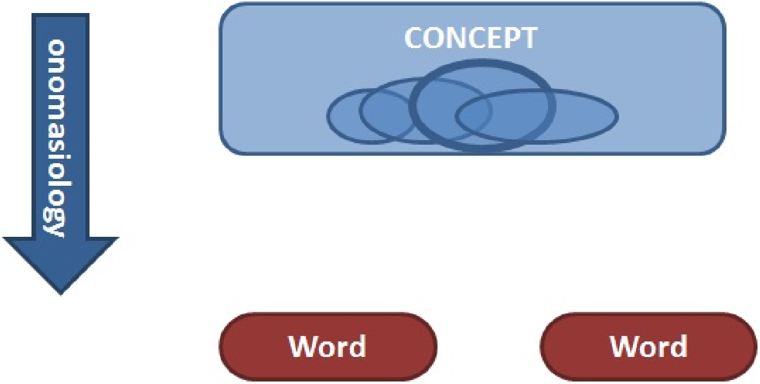
Linguistic Background

Structure of Lexical Variation (Geeraerts et al. 1994):



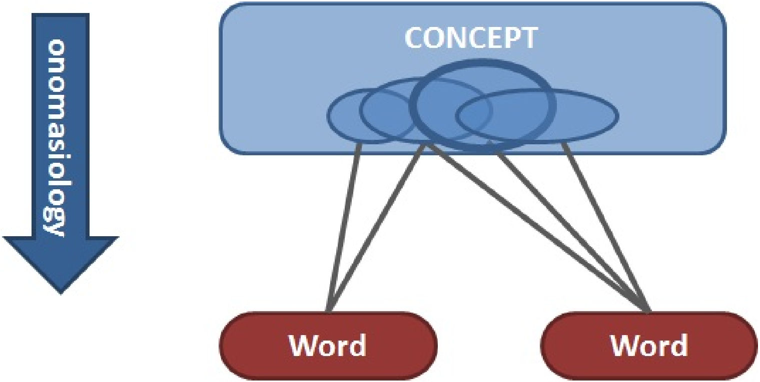
Linguistic Background

Structure of Lexical Variation (Geeraerts et al. 1994):
PROTOTYPE STRUCTURE:



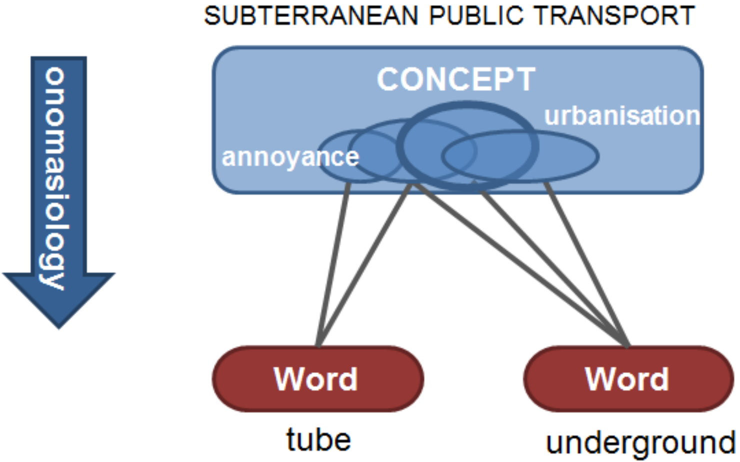
Linguistic Background

Structure of Lexical Variation (Geeraerts et al. 1994):
PROTOTYPE STRUCTURE:



Linguistic Background

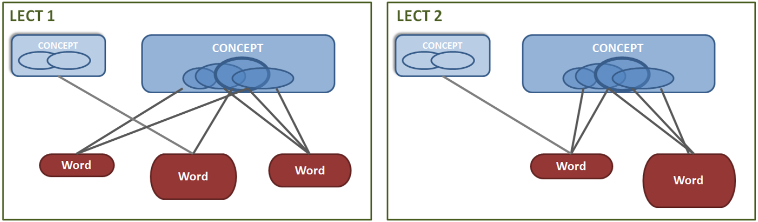
Structure of Lexical Variation (Geeraerts et al. 1994):
PROTOTYPE STRUCTURE:



Linguistic Background

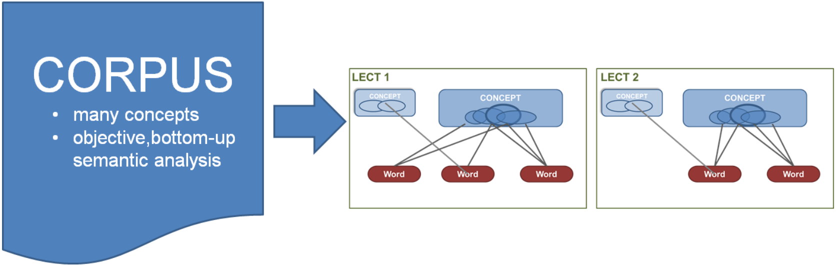
Structure of Lexical Variation (Geeraerts et al. 1994):

LECTAL VARIATION:



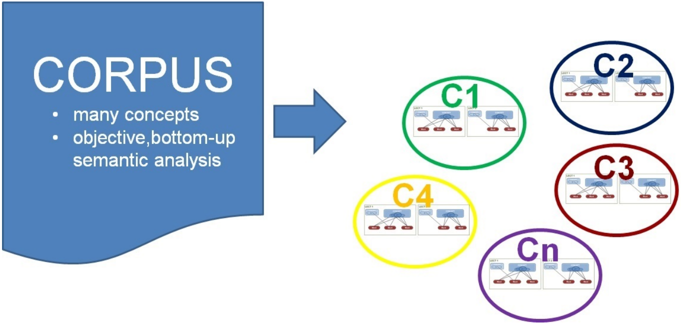
Linguistic Background

Structure of Lexical Variation (Geeraerts et al. 1994):
BASED ON BIG DATA:



Linguistic Background

Structure of Lexical Variation (Geeraerts et al. 1994):
BASED ON BIG DATA:



⇒ Automatic modelling of lexical semantics

Overview

1. Linguistic Background
2. Semantic Vector Spaces
3. Visual Analytics
4. Creating a 'gold standard' and cluster evalution.
5. Discussion and future work

2. Semantic Vector Spaces

Linguistic origin: Distributional Hypothesis

- "You shall know a word by the company it keeps" (Firth)
- a word's meaning can be induced from its co-occurring words
- long tradition of collocation studies in corpus linguistics

Semantic Vector Spaces in Computational Linguistics

- standard technique in statistical NLP for the large-scale automatic modeling of (lexical) semantics
- aka Vector Spaces Models, Distributional Semantic Models, Word Spaces,... (cf Turney & Pantel 2010 for overview)
- generalised, large-scale collocation analysis
- mainly used for automatic thesaurus extraction:
  $\Rightarrow$ words occurring in same contexts have similar meaning

Type-level SVS

Collect co-occurrence frequencies for a large part of the vocabulary and put them in a matrix

	transport	train	commute	ticket	scene	sugar	cream	now
subway	120	424	388	82	12	11	3	189
underground	154	401	376	99	305	20	1	123
coffee	5	8	18	4	1	72	102	152

Type-level SVS

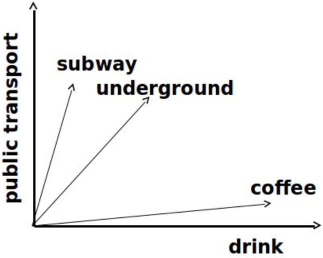
weight the raw frequencies by collocational strength (pmi)

	transport	train	commute	ticket	scene	sugar	milk	now
subway	5.3	7.9	6.5	4.0	0.8	0.6	0.0	0.0
underground	4.3	8.1	5.7	3.2	6.2	0.5	0.0	0.1
coffee	0.1	0.2	0.4	0.1	0.0	6.4	7.2	0.1

Type-level SVS

calculate word by word similarity matrix

	subway	underground	coffee
subway	1	.71	.08
underground	.71	1	.09
coffee	.08	.09	1



Token-level SVS

Make a vector for each occurrence of the variants

the teacher saw the dog chasing the cat

Token-level SVS

Make a vector for each occurrence of the variants

	3.2	4.3			0.8		7.1
	5.1	2.2			3.7		0.1
	0.2	3.5			2.3		0.3
	3.1	1.9			2.9		4.1
	4.7	0.2			1.3		3.1
	2.2	3.1			4.1		3.8
the	teacher	saw	the	dog	chasing	the	cat

Token-level SVS

Make a vector for each occurrence of the variants

					AVERAGE
3.2	4.3		0.8	7.1	3.9
5.1	2.2		3.7	0.2	2.8
0.2	3.5		2.3	0.3	1.6
3.1	1.9		2.9	4.1	3.0
4.7	0.2		1.4	3.1	2.3
2.2	3.1		4.1	3.8	3.3
teacher	saw	dog	chasing	cat	

Token-level SVS

Weighting

	3.2	4.3		0.8	7.1
	5.1	2.2		3.7	0.1
	0.2	3.5		2.3	0.3
	3.1	1.9		2.9	4.1
	4.7	0.2		1.3	3.1
	2.2	3.1		4.1	3.8
	teacher	saw	dog	chasing	cat
PMI weights	0.4	0.8		2.1	1.5

Context words are not equally informative for the meaning of dog.

Token-level SVS

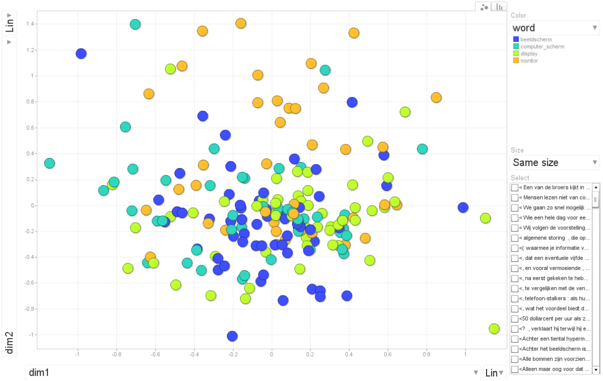
Weighted vectors

					WEIGHTED AVERAGE
3.2x0.4	4.3x0.8		0.8x2.1	7.1x1.5	4.3
5.1x0.4	2.2x0.8		3.7x2.1	0.2x1.5	3.0
0.2x0.4	3.5x0.8		2.3x2.1	0.3x1.5	2
3.1x0.4	1.9x0.8		2.9x2.1	4.1x1.5	3.8
4.7x0.4	0.2x0.8		1.4x2.1	3.1x1.5	2.4
2.2x0.4	3.1x0.8		4.1x2.1	3.8x1.5	4.4
teacher	saw	dog	chasing	cat	

Visual Analytics: Token clouds

Calculate similarity between all tokens

Version 1: use MDS and googlevis to plot interactively in 2D



Calibration problem

Semantic Vector Spaces, and especially token-level SVSs are parameter-rich.

Examples of parameters

- Bag-of-Words $\leftrightarrow$ Dependency Models
- Size of the context window for co-occurrences
- Size of the context window for weights
- Weighting scheme:
  Pointwise Mutual Information $\leftrightarrow$ Log-Likelihood Ratio
- Include $\leftrightarrow$ exclude highly-frequent (function words) words

Overview

1. Linguistic Background
2. Semantic Vector Spaces
3. Visual Analytics
4. Creating a 'gold standard' and cluster evalution.
5. Discussion and future work

3. Visual Analytics

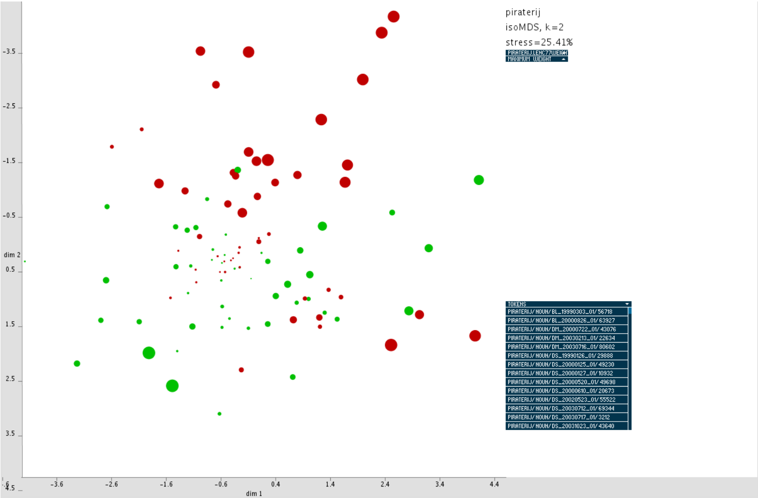
- Calibration could benefit from visual analytics of the different solutions.
- Using manually disambiguated data facilitates the visual evaluation as we can color–code the tokens for their different meanings.
- Misclassified tokens are quickly identified.
- We built our own, customisable tool to explore these token clouds.

3. Visual Analytics

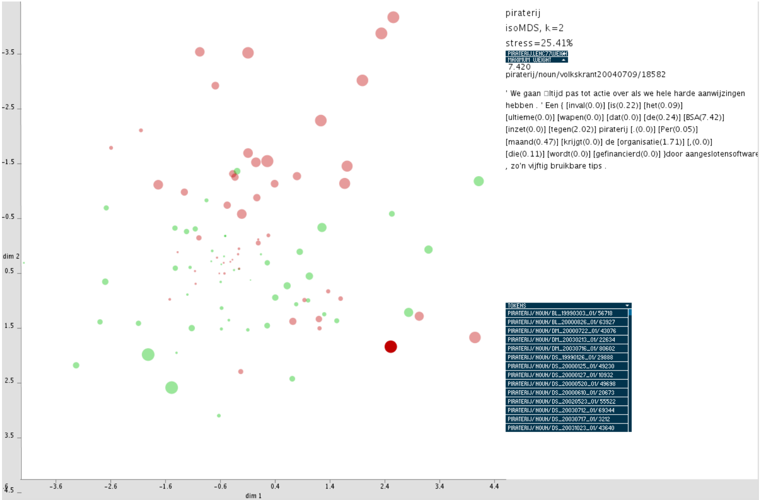
Dutch noun *piraterij*

- Data from large Dutch newspaper corpora
  Leuven News Corpus (LeNC): 1.3 billion words
  Twente News Corpus (TwNC): 500 million words
- Manually disambiguted data for the Dutch word type *piraterij* (piracy)
  piraterij$_1$: attack on ships
  piraterij$_2$: illegaly producing and selling products

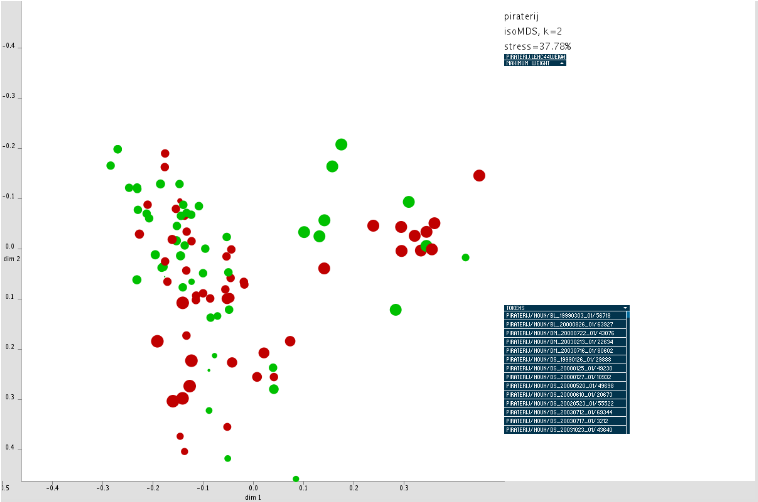
3. Visual Analytics



3. Visual Analytics



3. Visual Analytics



Overview

1. Linguistic Background
2. Semantic Vector Spaces
3. Visual Analytics
4. Creating a 'gold standard' and cluster evalution.
5. Discussion and future work

'Gold standard'

Manual effort

Selection of nouns from Algemeen Nederlands Woordenboek (ANW)

- Highly frequent in both BE and NL newspaper corpus.
- Examples that are not purily literary use.
- At least 2 core senses with a semantic relationship (betekenisbetrekking).

Manual disambiguation of random tokens until each sense has at least 50 occurrences.

'Gold standard'

ANW selection

- aanbieder (offerer)
- koper (buyer / copper)
- match
- motor (engine / motorcycle)
- parachute
- piraterij (piracy)
- pony
- prof
- scout
- therapeut (therapist)

'Gold standard'

ANW selection

- aanbieder (offerer)
- koper (buyer / copper)
- match
- motor (engine / motorcycle)
- parachute
- **piraterij** (piracy)
- pony
- prof
- **scout**
- **therapeut** (therapist)

4. Cluster evaluation

Aggregate cluster quality

- First proposed by McClain and Rao (1975) to evaluate clustering in marketing research.
- Speelman and Geeraerts (2009) proposed a similar measure for dialectometry.

$$\text{clusterqual: } \frac{S_W/N_W}{S_B/N_B}$$

S_W: within distances
N_W: number of distances between pairs
S_B: between distances
N_B: number of distances between pairs

4. Cluster evaluation

clusterqual properties

Due to its design:

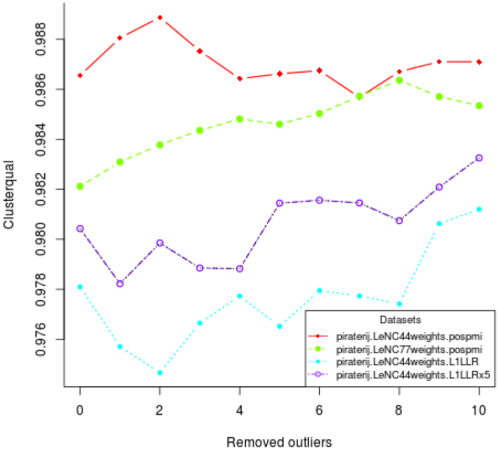
- clusterqual is sensitive to outliers.
- Unbalanced samples bias the result as our SemEval case study showed. (Wielfaert et al. 2013)

Solution:

- For each token, iteratively remove the n furthest tokens.
- Balance the sample over the different senses: 50 occurrences per sense.

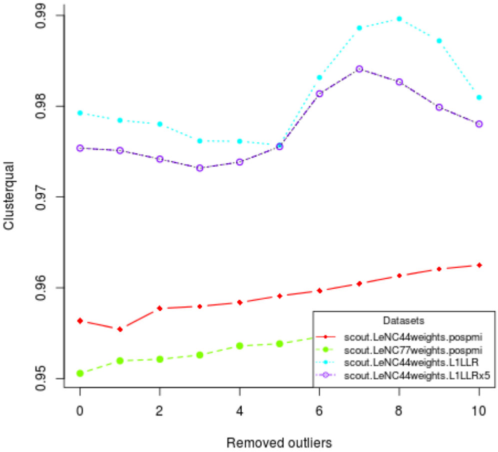
4. Cluster evaluation

piraterij



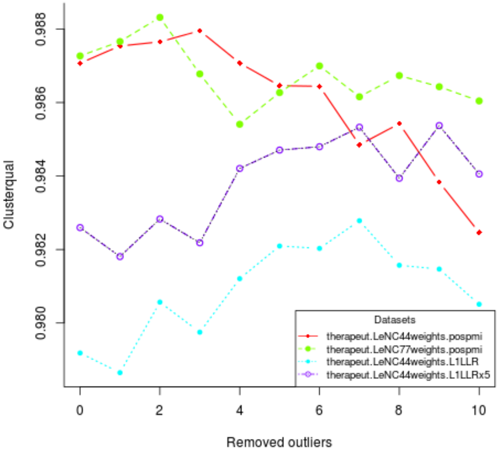
4. Cluster evaluation

scout



4. Cluster evalution

therapeut



Overview

1. Linguistic Background
2. Semantic Vector Spaces
3. Visual Analytics
4. Creating a 'gold standard' and cluster evalution.
5. Discussion and future work

5. Discussion and future work

'Gold standard' as a tool for parameter choice

- Controlled sample for different target words reduces the risk of overfitting.
- Finding one fits all parameter settings is probably impossible.

5. Discussion and future work

Extending the varied parameters

- Focus on weighting scheme of first-order co-occurrences, effect rather limited.
- Previous experiments: reducing noise largest improvement so far.
- Next step: remove function words and set low weights to virtually zero.

5. Discussion and future work

Other cluster quality indices

- clusterqual has its flaws
- Whole rang of other indices implemented in R *clusterCrit* package.

5. Discussion and future work

Fitting a model

- Number solutions grow quickly explodes when varying more parameters.
- Lapesa and Evert (2013) fitted a linear model on DSM parameters for 38800 solutions.

Purpose of the talk

THEORETICAL

- Study the structure of lexical variation: mapping of meaning onto lexemes in different varieties.
- Analyse how this structure is apparent in usage data

METHODOLOGICAL

- Semantic Vector Spaces as a method for the quantitative, large-scale, corpus-based analysis of lexical semantics
- Interactive Visualisation of distributional models as an exploratory, visual analytic tool for lexicology
- Creating a 'gold standard' and cluster evaluation.

For more information:

http://wwwling.arts.kuleuven.be/qlvl
thomas.wielfaert@arts.kuleuven.be
kris.heylen@arts.kuleuven.be
jocelyne.daems@arts.kuleuven.be
dirk.speelman@arts.kuleuven.be
dirk.geeraerts@arts.kuleuven.be