



Dutch Word Space Models: Context Selection and Evaluation Methods.

Yves Peirsman & Kris Heylen



KULeuven

Quantitative Lexicology and Variational Linguistics

Purpose

- Semantics in CL: using context to model word meaning
⇒ Word Space Models
- Compare models that use different definitions of context
- Evaluate the kind of semantic relations that are captured by these different models
⇒ more informed model choices for specific applications



Overview

1. Introduction
2. Context Definition
3. Results
4. Applications
5. Conclusions



Overview

1. Introduction

2. Context Definition

3. Results

4. Applications

5. Conclusions



1. Introduction

Word Space Models

Automatic identification of semantically related words

Distributional Hypothesis (Harris 1954, Firth 1957)

Words appearing in similar contexts will have similar meanings

Method

Based on a corpus, each target word is assigned a context vector, stating in which contexts the target word occurs and how often



1. Introduction

Target word in context

Eerder deze ochtend veroorzaakte een **ongeval** op de Brusselse Ring een kilometerslange file.

Context features in context vector

1x eerder, 1x ochtend, 1x veroorzaken, 1x Brusselse, 1x Ring

	eerder	ochtend	veroorzaken	Brussels	ring
ongeval	1	1	1	1	1



ongeval

0

auto

0

slachtoffer

0

vrachtwagen

0

file

0

gekwetst

0

suiker

0

melk

0

kop

	<i>auto</i>	<i>slachtoffer</i>	<i>vrachtwagen</i>	<i>file</i>	<i>gekwetst</i>	<i>suiker</i>	<i>melk</i>	<i>kop</i>
ongeval	0	0	1	0	1	0	0	0

vader raakte **gekwetst** bij een ongeval met een **vrachtwagen** op de

	<i>auto</i>	<i>slachtoffer</i>	<i>vrachtwagen</i>	<i>file</i>	<i>gekwetst</i>	<i>suiker</i>	<i>melk</i>	<i>kop</i>
ongeval	0	0	1	1	1	0	0	0

voor zeven uur veroorzaakte een ongeval een kilometerslange **file** richting Antwerpen

	<i>auto</i>	<i>slachtoffer</i>	<i>vrachtwagen</i>	<i>file</i>	<i>gekwetst</i>	<i>suiker</i>	<i>melk</i>	<i>kop</i>
ongeval	0	1	2	1	1	0	0	0

vrachtwagens waren betrokken bij het ongeval, dat meer dan tien *slachtoffers*

	<i>auto</i>	<i>slachtoffer</i>	<i>vrachtwagen</i>	<i>file</i>	<i>gekwetst</i>	<i>suiker</i>	<i>melk</i>	<i>kop</i>
ongeval	1	2	2	1	1	0	0	0

	<i>auto</i>	<i>slachtoffer</i>	<i>vrachtwagen</i>	<i>file</i>	<i>gekwetst</i>	<i>suiker</i>	<i>melk</i>	<i>kop</i>
ongeval	1	3	2	1	2	0	0	0

	<i>auto</i>	<i>slachtoffer</i>	<i>vrachtwagen</i>	<i>file</i>	<i>gekwetst</i>	<i>suiker</i>	<i>melk</i>	<i>kop</i>
ongeval	1	4	2	1	4	0	0	0

	<i>auto</i>	<i>slachtoffer</i>	<i>vrachtwagen</i>	<i>file</i>	<i>gekwetst</i>	<i>suiker</i>	<i>melk</i>	<i>kop</i>
ongeval	37	83	142	17	66	0	0	0

	<i>auto</i>	<i>slachtoffer</i>	<i>vrachtwagen</i>	<i>file</i>	<i>gekwetst</i>	<i>suiker</i>	<i>melk</i>	<i>kop</i>
ongeval	120	424	388	82	270	11	3	1

	<i>auto</i>	<i>slachtoffer</i>	<i>vrachtwagen</i>	<i>file</i>	<i>gekwetst</i>	<i>suiker</i>	<i>melk</i>	<i>kop</i>
ongeval	120	424	388	82	270	11	3	1
ongeluk	0	1	2	2	0	0	0	0

	<i>auto</i>	<i>slachtoffer</i>	<i>vrachtwagen</i>	<i>file</i>	<i>gekwetst</i>	<i>suiker</i>	<i>melk</i>	<i>kop</i>
ongeval	120	424	388	82	270	11	3	1
ongeluk	1	3	2	4	1	0	0	0

	<i>auto</i>	<i>slachtoffer</i>	<i>vrachtwagen</i>	<i>file</i>	<i>gekwetst</i>	<i>suiker</i>	<i>melk</i>	<i>kop</i>
ongeval	120	424	388	82	270	11	3	1
ongeluk	10	17	22	7	0	0	0	1

	<i>auto</i>	<i>slachtoffer</i>	<i>vrachtwagen</i>	<i>file</i>	<i>gekwetst</i>	<i>suiker</i>	<i>melk</i>	<i>kop</i>
ongeval	120	424	388	82	270	11	3	1
ongeluk	53	121	67	24	55	2	0	3

	<i>auto</i>	<i>slachtoffer</i>	<i>vrachtwagen</i>	<i>file</i>	<i>gekwetst</i>	<i>suiker</i>	<i>melk</i>	<i>kop</i>
ongeval	120	424	388	82	270	11	3	1
ongeluk	154	401	376	99	305	20	1	5

	<i>auto</i>	<i>slachtoffer</i>	<i>vrachtwagen</i>	<i>file</i>	<i>gekwetst</i>	<i>suiker</i>	<i>melk</i>	<i>kop</i>
ongeval	120	424	388	82	270	11	3	1
ongeluk	154	401	376	99	305	20	1	5
koffie	0	0	0	0	0	1	2	2

	<i>auto</i>	<i>slachtoffer</i>	<i>vrachtwagen</i>	<i>file</i>	<i>gekwetst</i>	<i>suiker</i>	<i>melk</i>	<i>kop</i>
ongeval	120	424	388	82	270	11	3	1
ongeluk	154	401	376	99	305	20	1	5
koffie	0	0	0	0	0	3	5	4

	<i>auto</i>	<i>slachtoffer</i>	<i>vrachtwagen</i>	<i>file</i>	<i>gekwetst</i>	<i>suiker</i>	<i>melk</i>	<i>kop</i>
ongeval	120	424	388	82	270	11	3	1
ongeluk	154	401	376	99	305	20	1	5
koffie	0	0	2	0	0	16	24	21

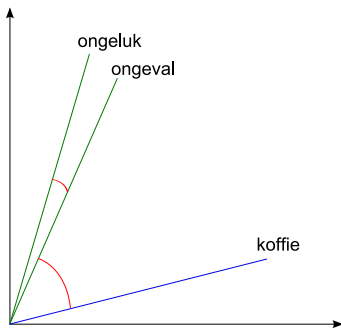
	<i>auto</i>	<i>slachtoffer</i>	<i>vrachtwagen</i>	<i>file</i>	<i>gekwetst</i>	<i>suiker</i>	<i>melk</i>	<i>kop</i>
ongeval	120	424	388	82	270	11	3	1
ongeluk	154	401	376	99	305	20	1	5
koffie	3	5	11	1	0	55	76	64

	<i>auto</i>	<i>slachtoffer</i>	<i>vrachtwagen</i>	<i>file</i>	<i>gekwetst</i>	<i>suiker</i>	<i>melk</i>	<i>kop</i>
ongeval	120	424	388	82	270	11	3	1
ongeluk	154	401	376	99	305	20	1	5
koffie	5	8	18	4	1	72	102	93

1. Introduction

Semantic distance

- vectors projected in context feature space: wordspaces
- cosine of angle between vectors as semantic similarity measure



Overview

1. Introduction
2. Context Definition
3. Results
4. Applications
5. Conclusions

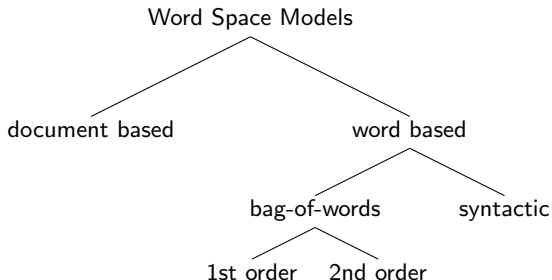


2. Context Definition

Word Space Models come in many flavours

The main difference between the models lies in how they define context

Family of models



2. Context Definition

document based models

- context = stretch of text in which target word occurs
- 2 words are related when they often co-occur in text
- Landauer & Dumais 1997: Latent Semantic Analysis

word based models

- context = context words around the target word
- 2 words are related when they co-occur with the same context words, but not necessarily with each other



	<i>DOC.1</i>	<i>DOC.2</i>	<i>DOC.3</i>	<i>DOC.4</i>	<i>DOC.5</i>	<i>DOC.6</i>	<i>DOC.7</i>	<i>DOC.8</i>
ongeval	23	12	14	24	8	0	0	0
ongeluk	16	9	11	18	17	20	0	1
koffie	0	0	0	0	0	14	12	15
<i>auto</i>		<i>slachtoffer</i>	<i>vrachtwagen</i>	<i>file</i>	<i>gekwetst</i>	<i>suiker</i>	<i>melk</i>	<i>kop</i>
ongeval	120	424	388	82	270	11	3	1
ongeluk	154	401	376	99	305	20	1	5
koffie	5	8	18	4	1	72	102	93

2. Context Definition

Within word based models:

bag-of-words

- context words in window of n words left and right of target word
- a bag of unstructured context features

syntactic features

- context words in specific syntactic relation with target word
- takes clause structure into account
- Lin 1998, Padó & Lapata 2007



De kwispelende hond blafte naar de postbode op de fiets

	<i>kwispelend</i>	<i>hond</i>	<i>blaffen</i>	<i>postbode</i>	<i>fiets</i>
<i>hond</i>	1	0	1	1	1
<i>postbode</i>	1	1	1	0	1

De kwispelende hond blafte naar de postbode op de fiets

	<i>subj.blaffen</i>	<i>+adj.kwispelend</i>	<i>PC.blaffen.naar</i>	<i>+PP.op.fiets</i>
hond	1	1	0	0
postbode	0	0	1	1

2. Context Definition

Within the bag-of-words models:

1st order co-occurrences

- context = words in immediate proximity to the target
- Levy & Bullinaria 2001

2nd order co-occurrences

- context = context words of context words of target
- can generalise over semantically related context words
- Schütze 1998

NB syntactic models are also 1st order models



's Morgens veroorzaakte een [ongeval](#) een file

's Morgens veroorzaakte een **ongeval** een file

Hij kan 's **morgens** zijn **bed** niet uit

Neem 's **morgens** **tijd** voor een **ontbijt**

's **Morgens** gaat de **wekker** altijd te **vroeg** af

Roken veroorzaakt kanker
CO2 veroorzaakt klimaatopwarming
De sneeuw veroorzaakt problemen

's Morgens veroorzaakte een ongeval een file

Hij kan 's morgens zijn bed niet uit
Neem 's morgens tijd voor een ontbijt
's Morgens gaat de wekker altijd te vroeg af

Roken veroorzaakt **kanker**
 CO2 veroorzaakt **klimaatopwarming**
 De **sneeuw** veroorzaakt **problemen**

's Morgens veroorzaakte een **ongeval** een **file**

Hij kan 's morgens zijn **bed** niet uit
 Neem 's morgens **tijd** voor een **ontbijt**
 's Morgens gaat de **wekker** altijd te **vroeg** af

Ze **stond uren** in de **file**
 Op de **ringweg staat** een **file**
 Met de **auto** sta je sowieso in de **file**

Roken veroorzaakt **kanker**
 CO2 veroorzaakt **klimaatopwarming**
 De **sneeuw** veroorzaakt **problemen**

's Morgens veroorzaakte een **ongeval** een file

Hij kan 's morgens zijn **bed** niet uit
 Neem 's morgens **tijd** voor een **ontbijt**
 's Morgens gaat de **wekker** altijd te **vroeg** af

Ze **stond uren** in de file
 Op de **ringweg staat** een file
 Met de **auto** sta je sowieso in de file

Roken veroorzaakt **kanker**
 CO2 veroorzaakt **klimaatopwarming**
 De **sneeuw** veroorzaakt **problemen**

's Ochtends veroorzaakte een **ongeval** een file

Hij kan 's ochtends zijn **bed** niet uit
 Neem 's ochtends **tijd** voor een **ontbijt**
 's ochtends gaat de **wekker** altijd te **vroeg** af

Ze **stond uren** in de file
 Op de **ringweg staat** een file
 Met de **auto** sta je sowieso in de file

2. Context Definition

Problems

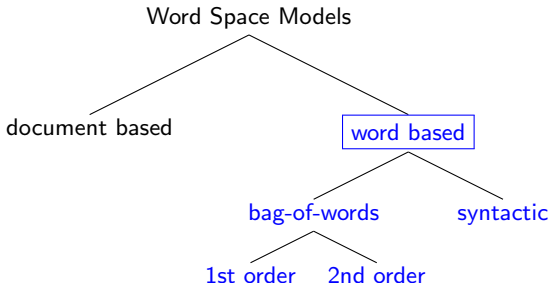
- “Comparisons between the two types of models have been few and far between in the literature.” (Padó & Lapata 2007)
- What kind of semantic similarity do the models capture?
- Sahlgren 2006: document based \Rightarrow syntagmatic relations;
word-based \Rightarrow paradigmatic relations
- Within word based models: no systematic analysis of the specific semantic relations they are supposed to capture...
- Crucial in choosing the model that is best suited for a specific application (QA, WSD, MT,...)



2. Context Definition

Research goals

- Compare different word-based models on the same data
- Context definition ranging on a continuum from strict to loose (syntactic 1st order > bag-of-words 1st order > bag-of-words 2nd order)
- Analyse the type of semantic relations found by these models.



Overview

1. Introduction
2. Context Definition
- 3. Results**
4. Applications
5. Conclusions



3.1 Experiments

Investigated models

- syntactic model
- first-order bag-of-word models with context sizes 1, 3 and 5
- second-order bag-of-word models with context sizes 1, 3 and 5

Parameters

- data: Twente Nieuws Corpus
- 2000 dimensions: the most frequent features in the corpus
- cut-off at frequency 5
- point-wise mutual information
- similarity calculated with cosine



3.2 Semantic relations

1. Semantic similarity
 - 1.1 Definition
 - 1.2 Gold Standard
 - 1.3 Results
2. Semantic relatedness
 - 2.1 Definition
 - 2.2 Gold Standard
 - 2.3 Results
3. Clusters of models

3.2.1 Semantic similarity: Definition

Definition

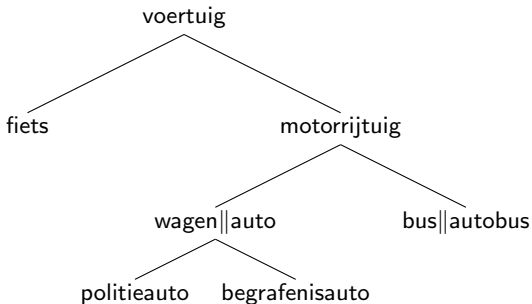
Two words are semantically similar when there is a relationship of similarity between the concepts they refer to.

Examples

- synonyms: jeansbroek – spijkerbroek, auto – wagen
- hyp(er)onyms: vogel – roodborstje, boek – roman
- co-hyponyms: roodborstje – merel, roman – gedichtenbundel

3.2.1 Semantic similarity: Gold Standard

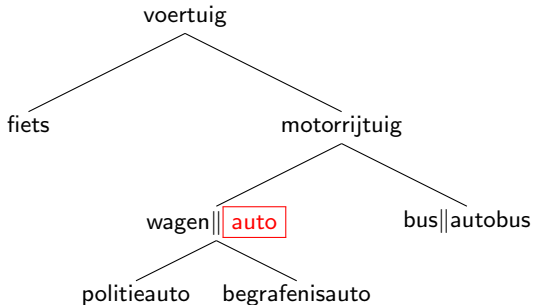
Dutch EuroWordNet



3.2.1 Semantic similarity: Gold Standard

Dutch EuroWordNet

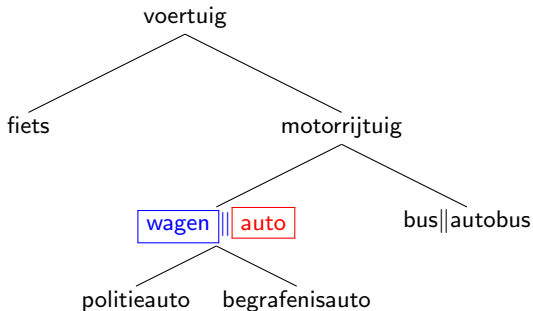
target word



3.2.1 Semantic similarity: Gold Standard

Dutch EuroWordNet

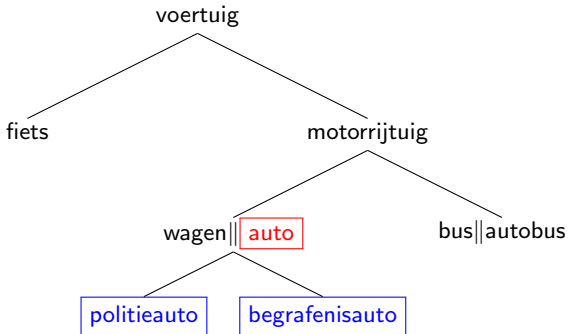
synonyms



3.2.1 Semantic similarity: Gold Standard

Dutch EuroWordNet

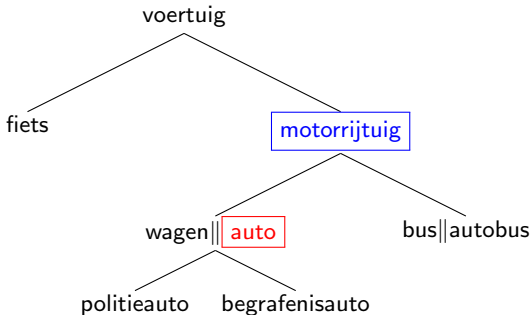
hyponyms



3.2.1 Semantic similarity: Gold Standard

Dutch EuroWordNet

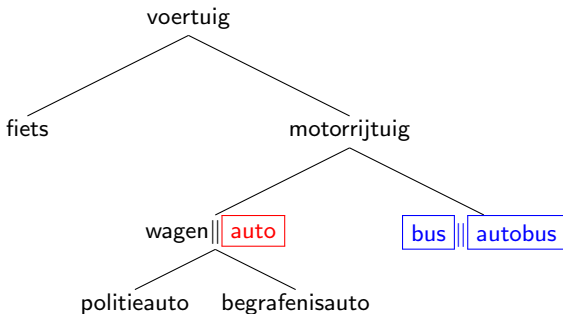
hypernyms



3.2.1 Semantic similarity: Gold Standard

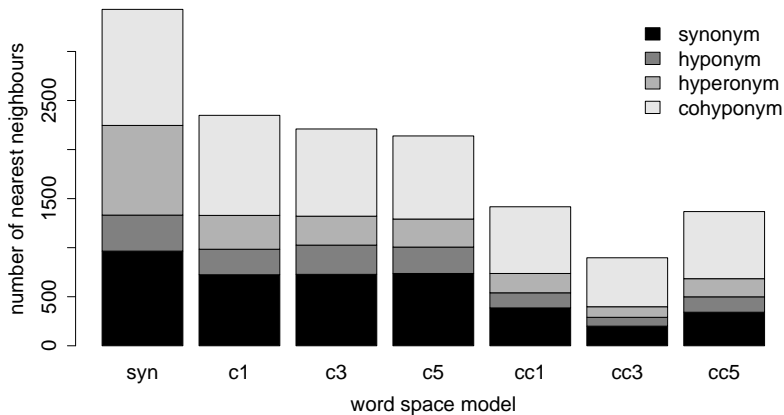
Dutch EuroWordNet

co-hyponyms



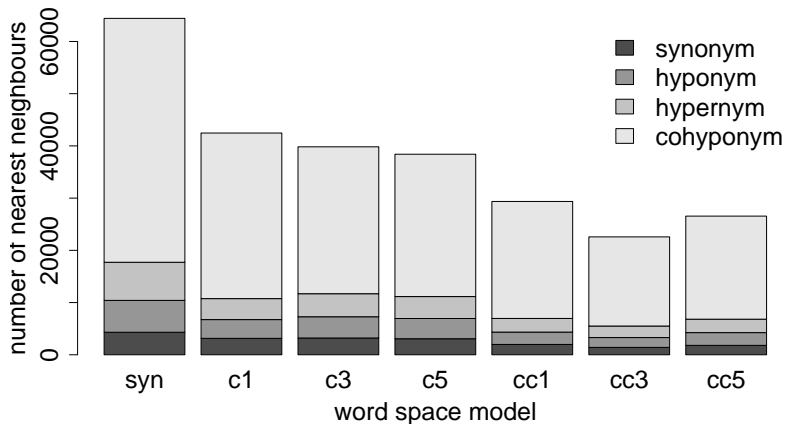
3.2.1 Semantic similarity: Results

1 neighbour

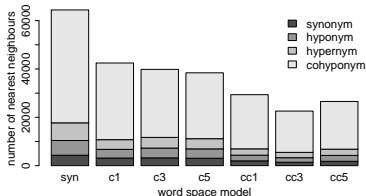
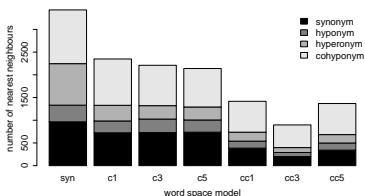


3.2.1 Semantische similarity: Results

100 neighbours



3.2.1 Semantic similarity: Results



- Syntactic model > first-order bag-of-words > second-order bag-of-words
- The “stricter” the definition of context, the more semantically similar words.
- The larger the context, the fewer co-hyponyms.



3.2.2 Semantic relatedness: Definition

Definition

The concepts to which the two words refer are part of the same *frame*.

- meronyms: *auto* — *stuur*
- locations: *vogel* — *nest*
- etc.: *lucifer* — *vuur*

Semantic similarity is a subtype of semantic relatedness.

3.2.2 Semantic relatedness: Gold Standard

Human associations

Experiment where participants were asked to give three associations for each *cue word*.

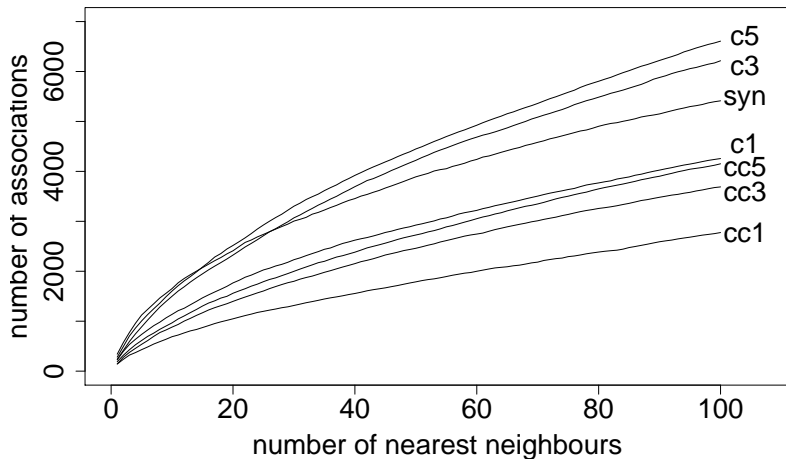
- aardappel: eten (22x), friet (14x), boer (5x), ...
- advocaat: rechtbank (21x), toga (17x), geld (15x), ...

Test set

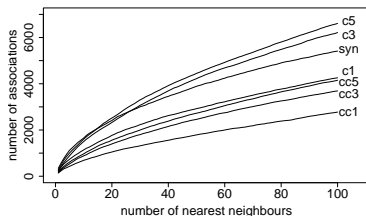
- Cue words and associations that belong to the 10.000 most frequent nouns (+30.000)
- Only 8% of the associations are part of the EuroWordNet environment of the cue word.



3.2.2 Semantic relatedness: Results



3.2.2 Semantic relatedness: Results

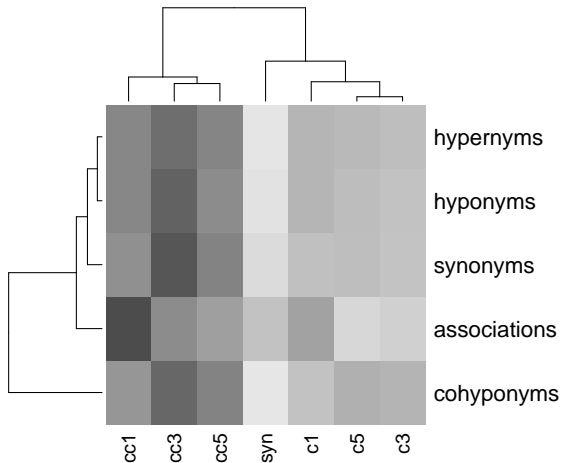


- With a small number of neighbours, the syntactic model scores best.
- Later the first-order bag-of-words models with context sizes 3 and 5 climb higher.



3.3 Clusters of models

Absolute frequency of the various relations



3.3 Clusters of models

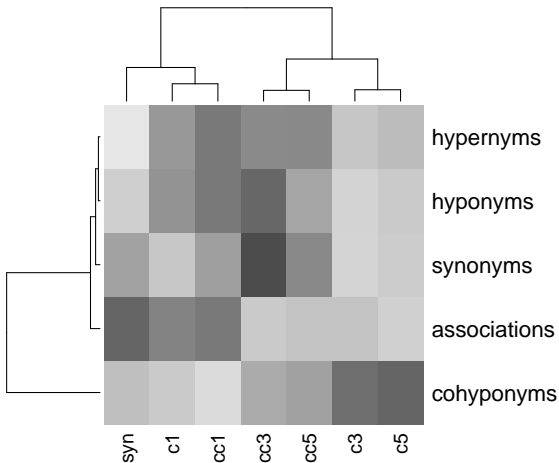
Success of models

- First-order models are clearly separated from second-order models.
- Syntactic model scores better for semantic similarity.
- Bag-of-word models with context sizes 3 and 5 better for associations.



3.3 Clusters of models

Relative frequency of the various relations



3.3 Clusters of models

Semantic preference of the models

- Small contexts have larger relative frequencies of co-hyponyms.
- Large contexts have larger relative frequencies of associations.
- First-order contexts have larger relative frequencies of synonyms, hyponyms and hypernyms.



Overview

1. Introduction
2. Context Definition
3. Results
4. Applications
5. Conclusions



4. Applications

Computational Linguistics

- Thesaurus extraction
- Word Sense Disambiguation
- Question Answering
- Anaphor resolution
- ...

Lexicology

- Usage Based Semantics
- Lexical variation research (Lectometry)



4. Applications

Profile based measurement of lexical variation

Do B and NL use the same lexemes to refer to a given concept?

- define a set of synonyms = profile
- collect all instances from 2 corpora (B vs NL) + disambiguate
- Calculate relative frequency of synonyms in 2 corpora
- Overlap in relative frequency = uniformity measure

	B	NL	overlap
jeans	85	30	30
spijkerbroek	15	70	15
			45

4. Applications

Previous Research

- sports and clothing terms: Geeraerts et al. 1999, 2003
- onomasiological, avoids thematic bias
- time consuming, not readily scalable

sem·metrix project

- tools for large scale lexical variation research
- automatic generation of synonym sets (profiles)
- Word Sense Disambiguation for profile membership



Overview

1. Introduction

2. Context Definition

3. Results

4. Applications

5. Conclusions



5. Conclusions

- Word Space Models with different context definitions
- Evaluation of 7 models with context definition ranging from strict to loose
- syntactic models are better at capturing specific semantic relations
- 1st order bag-of-word models with larger context windows approximate humans' intuitive associations better
- These insights allow for a more informed choice of Word Space Model for a specific application





For more information:

<http://wwling.arts.kuleuven.be/qlvl>

yves.peirsman@arts.kuleuven.be

kris.heylen@arts.kuleuven.be

