Modelling Word Similarity

An Evaluation of Automatic Synonymy Extraction Algorithms

Kris Heylen, Yves Peirsman, Dirk Geeraerts, Dirk Speelman

KULeuven
Quantitative Lexicology and Variational Linguistics

Purpose

- Use Word Space Models to find synonyms
- Compare models with different definitions of context
- Evaluate whether these models do equally well for all words:
  frequent and infrequent, specific and general terms, abstract and concrete

  ⇒ more informed model choices for specific applications

Overview

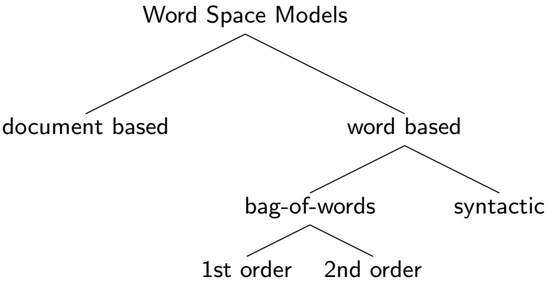
Overview

Introduction

Words Space or Distributional Models

- Words appearing in similar contexts have similar meanings
- Word meaning is modelled as a vector of context features
- Semantic similarity is measured as context vector similarity

Different context definitions:



Introduction

document based models

- context = text in which target word occurs (e.g. documents)
- 2 words are related when they often co-occur in documents
- Landauer & Dumais 1997: Latent Semantic Analysis

word based models

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

Introduction

Within word based models:

bag-of-words

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

syntactic features

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

Introduction

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

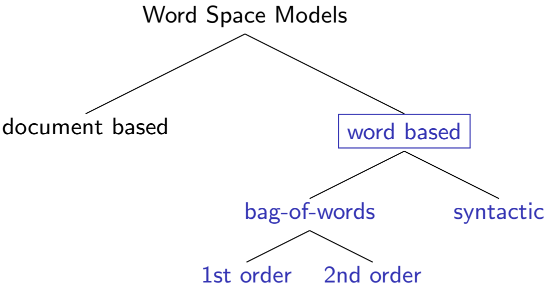
Introduction

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 these models actually capture?
- Do they work equally well for all types of target words?
- Crucial in choosing the model that is best suited for a specific application (QA, WSD, IR,...)

Research goals

- Compare word-based models with different context definitons on the same data
- Analyse the type of semantic relations found
- Evaluate whether retrieval works equally well for different classes of target words



Overview

1. Introduction
2. Experimental setup
3. Evaluation scheme
4. Influence of word properties
5. Conclusions

Experimental setup

Three Word Space Models for Dutch

- first order bag of words
- second order bag of words
- syntactic (dependency-based)

Variation on 2 parameters

- context type: mere co-occurrence vs syntactic dependency
- order: 1st order vs 2nd order co-occurrences

Experimental setup: Context type

Bag of words

mere co-occurrence: words that appear at least 5 times in a context window of n words around the target word w.

Syntactic contexts

dependency relations: subject, direct object, prepositional complement, adverbial prepositional phrase, adjectival modification, PP postmodification, apposition, coordination

Experimental setup: Order

1st order

words that occur in immediate proximity to the target word w.

2nd order

words that co-occur with the 1st order co-occurrence of the target word w.

$\Rightarrow$ *Only varied for BoW models, although, in principle, 2nd order syntactic relations possible as well*

Experimental setup: other parameters

- Window size (b-o-w): 3 words left and right
- Dimensionality: fixed at 4000 most frequent features,
  - cut-off of 5 (bag-of-words)
  - experiments with Random Indexing (Peirsman & Heylen 2007)
- Weighting scheme: point-wise mutual information index
- Similarity measure: cosine between vectors
- Data: Twente Nieuws Corpus, 300M words of newspaper text, parsed with Alpino (van Noord 2006)
- Test set: 10,000 most frequent nouns

Overview

1. Introduction
2. Experimental setup
3. Evaluation scheme
4. Influence of word properties
5. Conclusions

Evaluation Scheme

Evaluated Output

- for each of the 10.000 target words, the semantically most similar word was retrieved = Nearest Neighbour (NN)
- by each of the three models (1^o bow, 2^o bow, dependency)

Evaluation Criteria

Gold Standard Dutch EuroWordNet (EWN) (even though...)

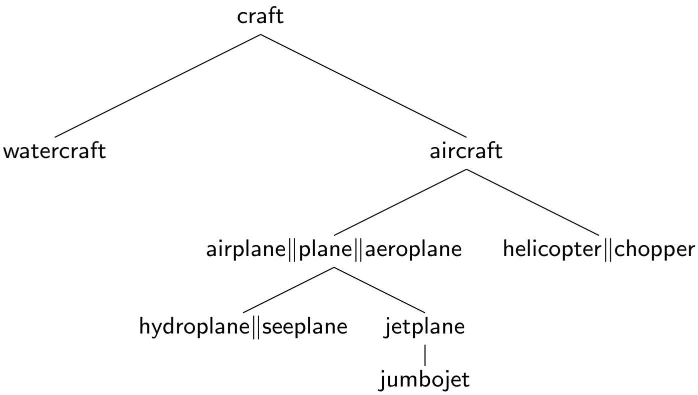
criterium 1 average Wu & Palmer score of NNs

criterium 2 % syno-, hypo-, hyper- en cohyponyms among NNs

NB: only pairs in EWN (syn 7479, 1^obow 6776, 2^obow 6727)

Evaluation Scheme

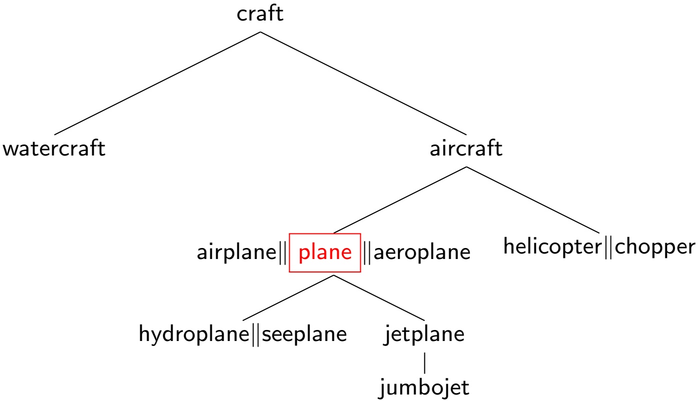
Definition of semantic relationships



Evaluation Scheme

Definition of semantic relationships

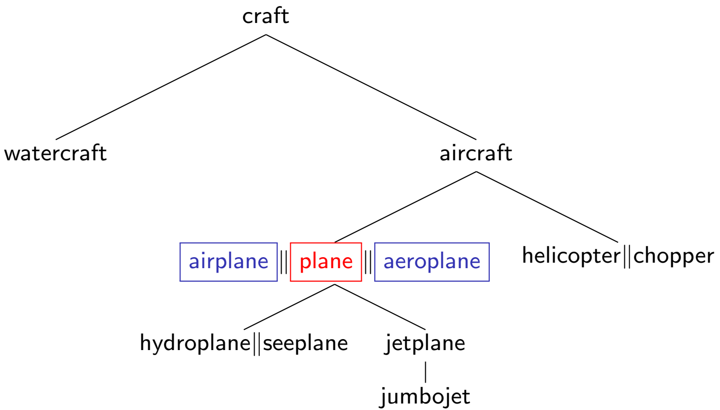
target word



Evaluation Scheme

Definition of semantic relationships

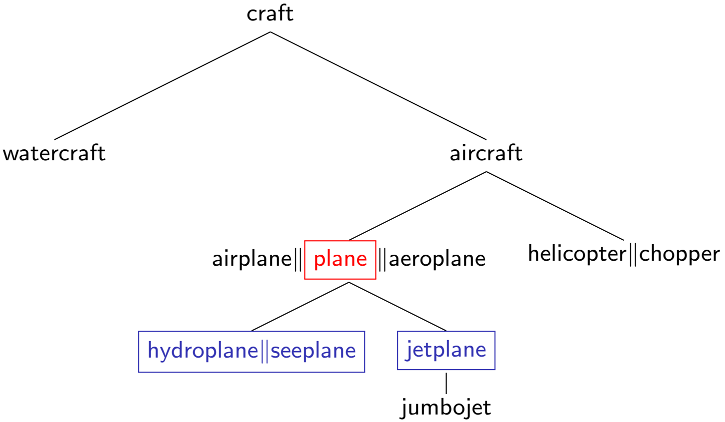
synonyms



Evaluation Scheme

Definition of semantic relationships

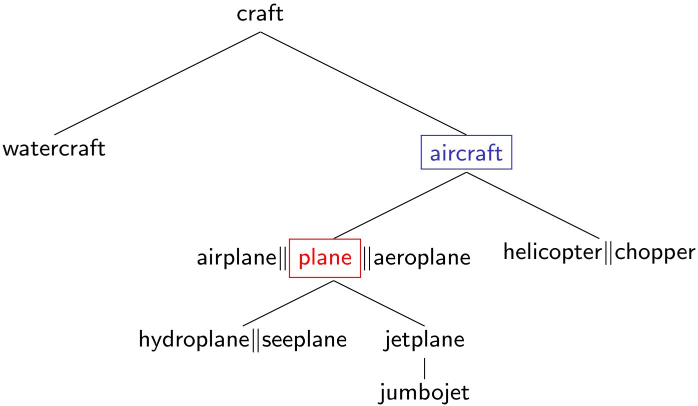
hyponyms



Evaluation Scheme

Definition of semantic relationships

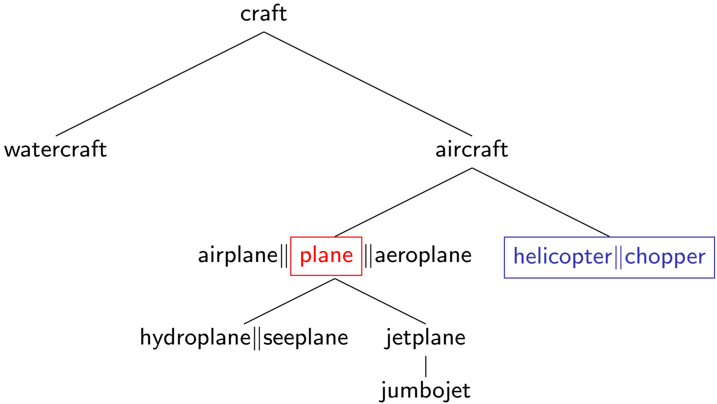
hypernyms



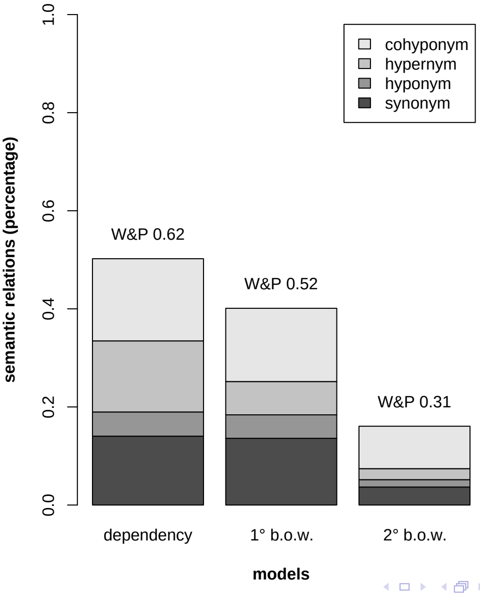
Evaluation Scheme

Definition of semantic relationships

co-hyponyms



Overall performance (Peirsman, Heylen & Speelman 2008)



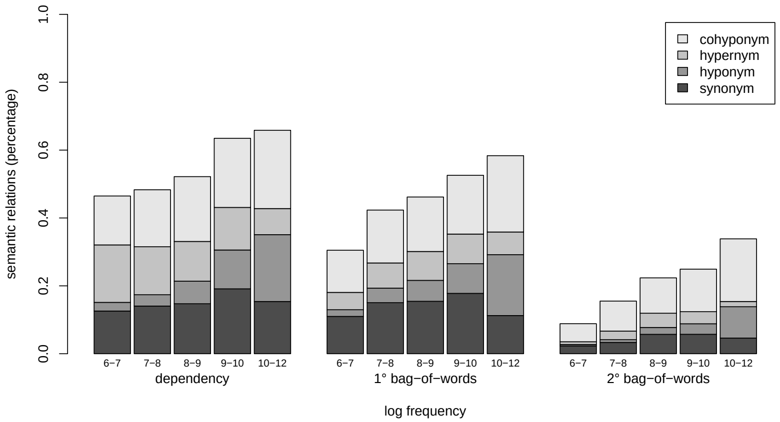
Overview

Results: Influence of word properties

- **Up to now:** no differentiation between target words
- **But:** Can synonyms be equally well retrieved for all classes of target words?
- **Question:** Do the linguistic properties of target words influence the perfomance of the models?
- Three properties:
  1. Frequency
  2. Semantic specificity
  3. Semantic class

Influence of Frequency

natural log of target word frequency in our corpus

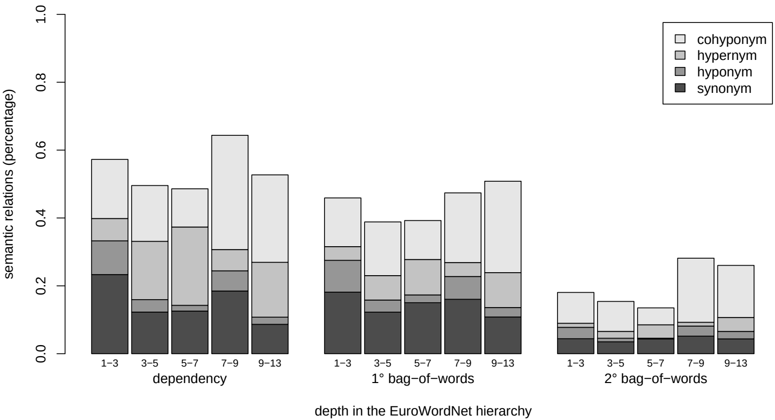


Influence of Frequency

- higher frequency $\Rightarrow$ more relations (synon. & hypon.)
- stronger effect for weak 2^o bow model
- possible explanations:
  - technical reason: more data for frequent words
  - more frequent words are more likely to have synonyms

Influence of Semantic Specificity

Depth of target word in WordNet hierarchy

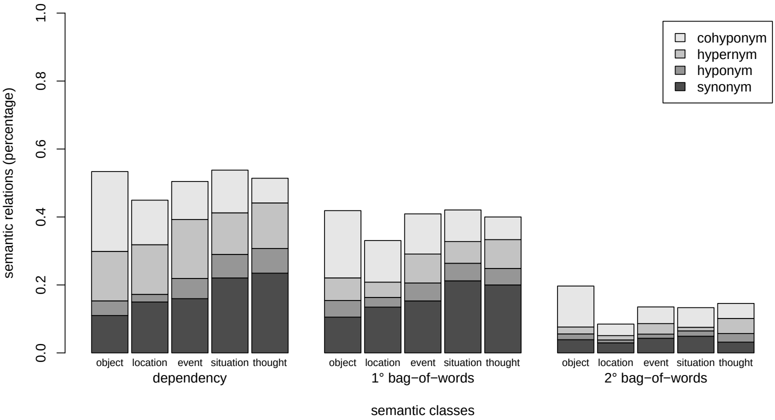


Influence of Semantic Specificity

- No clear (linear) effect
- more synonyms for unspecific and intermediately specific terms
- intermediates mainly person nouns (teacher, thief, villain)
- possible explanations
    - Base level categories?
    - Granularity variance in EWN

Influence of Semantic Class

the but 1 highest ancestor in WordNet (5 out of 41):
object, location, event, situation, thought



Influence of Semantic Class

- number of related NNs remains constant
- significantly more synonyms for *thoughts* than for *objects*
- cline concrete-abstract: more synonyms for abstract words
- possible explanations
  - better represented in newspaper data
  - fuzzyness of abstract categories
  - more readily put in same synset in EWN

Overview

1. Introduction
2. Experimental setup
3. Evaluation scheme
4. Influence of word properties
5. Conclusions

Conclusions

Influence of target word properties on the perfomance of Word Space Models for Dutch

- tighter semantic relations for high frequency words
- no clear effect of semantic specificity
- more synonyms retrieved for abstract semantic classes
- similar effects for 1^o, 2^o bow and syntactic model
- syntactic model best performing for any subclass of words

Future work

- find out WHY these properties have an effect
- words from specific topical domains

For more information:

http://wwwling.arts.kuleuven.be/qlvl

kris.heylen@arts.kuleuven.be

yves.peirsman@arts.kuleuven.be