



Automatic Thesaurus Extraction: Comparing context models.

Kris Heylen & Yves Peirsman



KULeuven

Quantitative Lexicology and Variational Linguistics

Purpose

- Find semantically similar words via the similar contexts they appear in
- Compare models that use different definitions of context
- Evaluate the kind of semantic similarity that is captured by these different models



Overview

1. Introduction
2. Experimental setup
3. Evaluation
4. Results
5. Conclusions



Overzicht

1. Introduction

2. Experimental setup

3. Evaluation

4. Results

5. Conclusions



1. Introduction

Distributional Hypothesis (Harris 1954)

Words that appear in similar contexts have similar meanings.

How should *context* be defined?

- bag of words
 - first order or second order
 - different window sizes
- syntactic relations
 - types of dependency relations
 - specificity of the dependency information
 - length of the dependency paths



Overzicht

1. Introduction
2. Experimental setup
3. Evaluation
4. Results
5. Conclusions



2. Experimental setup

Variations on 3 parameters

- **context type:** mere co-occurrence vs syntactic dependency
- **order:** 1st order vs 2nd order co-occurrences
- **context size:** size of the co-occurrence window



2. 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

Hypothesis

Syntactic contexts capture tighter semantic relations than co-occurrences



2. Experimental setup: Order

1st order

words that co-occur with the target word w .

2nd order

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

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

Hypothesis

1st order contexts capture tighter semantic relations than 2nd order contexts



2. Experimental setup: Context size

Window size

n varies between 1 and 20 (1, 3, 5, 7, 10, 15, 20)

⇒ *Only varied for 1st order BoW models. In principle, also possible for 2nd order BoW models and syntactic models, where context size corresponds to length of the dependency path.*

Hypothesis

Smaller windows capture tighter semantic relations, but may suffer from data sparseness.



2. Experimental setup: other parameters

- **Dimensionality:** fixed at 4000 most frequent features
 - 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



Overzicht

1. Introduction
2. Experimental setup
- 3. Evaluation**
4. Results
5. Conclusions



3. Evaluation method 1: Overlap

How similar are the results of the different models?

For each target word:

- take the n nearest neighbours according to model 1

- take the n nearest neighbours according to model 2

Calculate the overlap

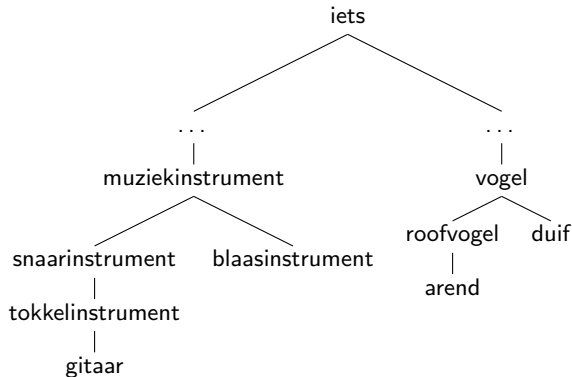
Calculate the average overlap for all target words



3. Evaluation method 2: Wu and Palmer

What is the overall quality of our models?

EuroWordNet (Vossen 1997)



3. Evaluation method 2: Wu and Palmer

For each target word:

take the n most nearest neighbours according to the model



3. Evaluation method 2: Wu and Palmer

For each target word:

take the n most nearest neighbours according to the model

For each nearest neighbour:

if it occurs in EuroWordNet:

calculate the Wu & Palmer similarity to the target word

else:

ignore



3. Evaluation method 2: Wu and Palmer

For each target word:

- take the n most nearest neighbours according to the model

- For each nearest neighbour:

 - if it occurs in EuroWordNet:

 - calculate the Wu & Palmer similarity to the target word

 - else:

 - ignore

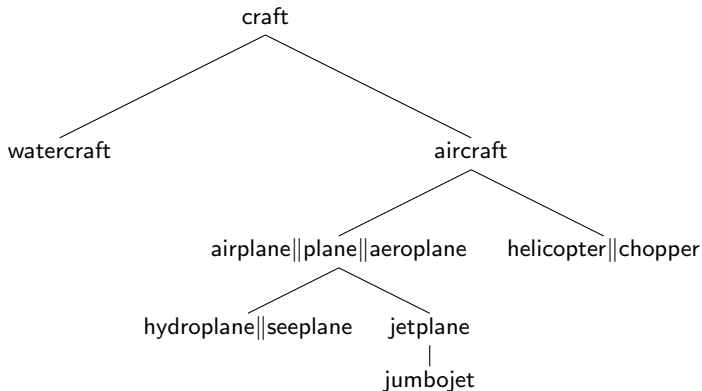
Calculate the average Wu & Palmer similarity for the target

Take the average over all targets



3. Evaluation method 3: semantic relations

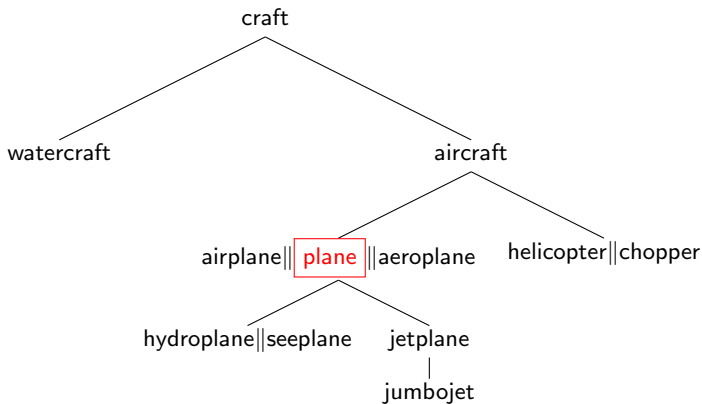
What semantic relationships do our models find?



3. Evaluation method 3: semantic relations

What semantic relationships do our models find?

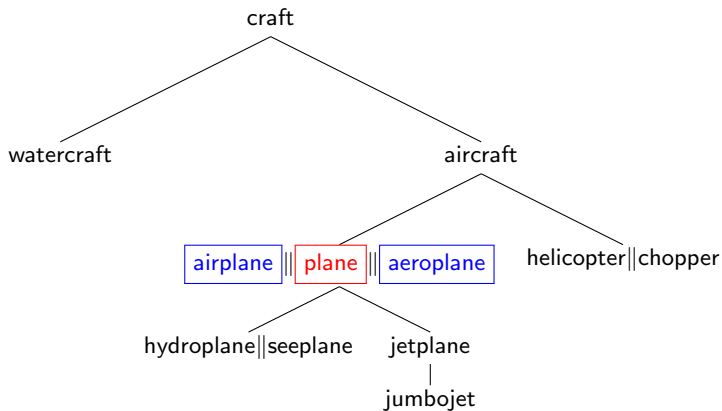
target word



3. Evaluation method 3: semantic relations

What semantic relationships do our models find?

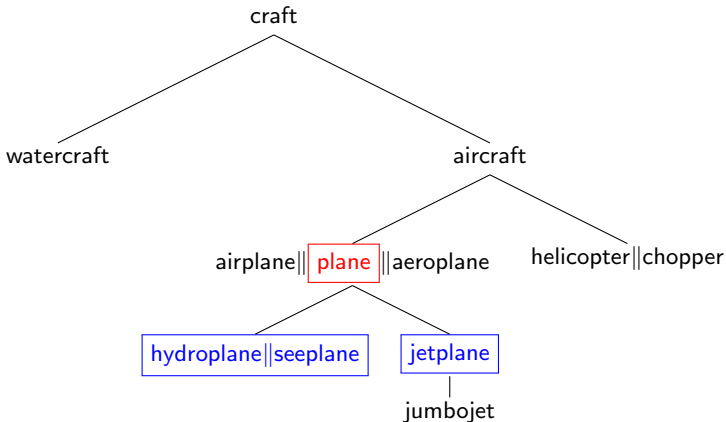
synonyms



3. Evaluation method 3: semantic relations

What semantic relationships do our models find?

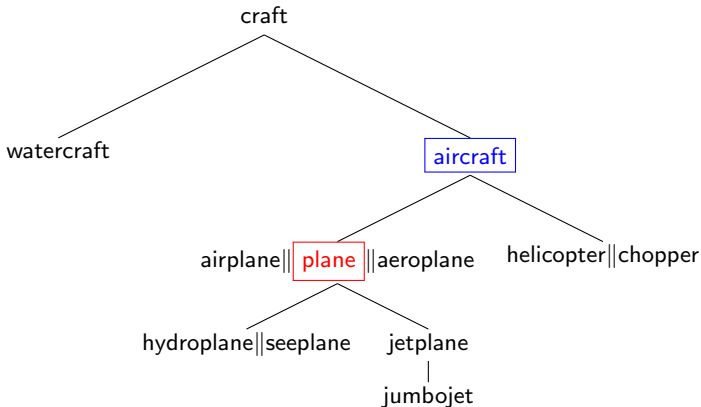
hyponyms



3. Evaluation method 3: semantic relations

What semantic relationships do our models find?

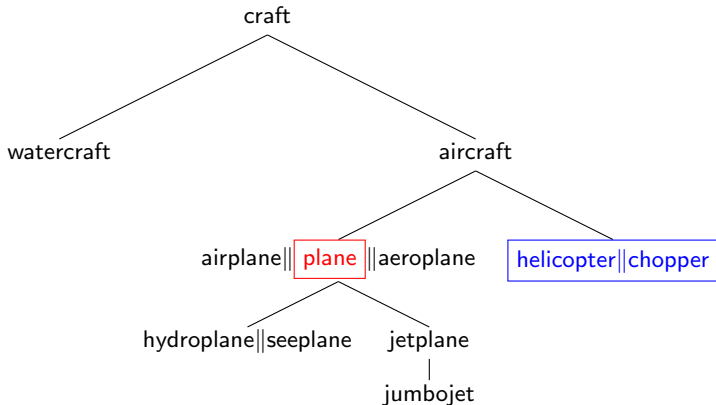
hypernyms



3. Evaluation method 3: semantic relations

What semantic relationships do our models find?

co-hyponyms

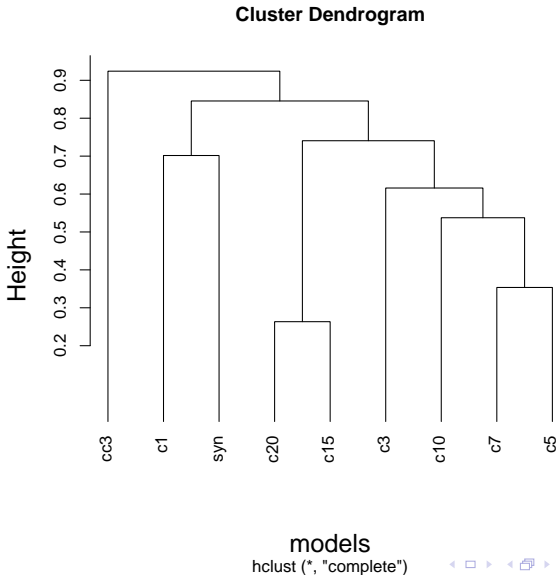


Overzicht

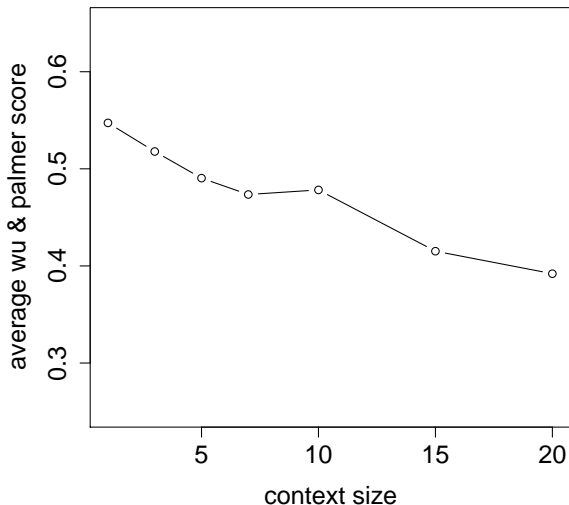
1. Introduction
2. Experimental setup
3. Evaluation
4. Results
5. Conclusions



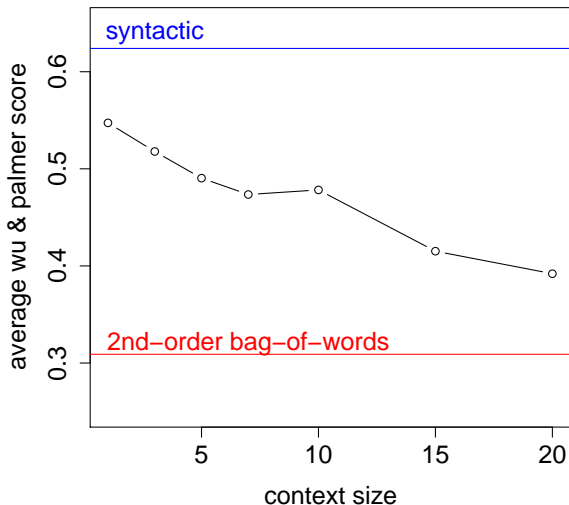
4. Results: how similar are the results?



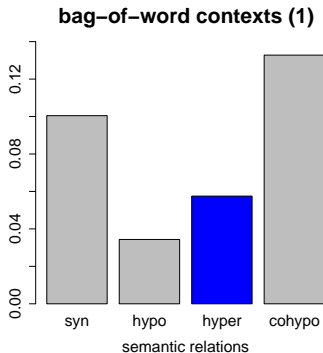
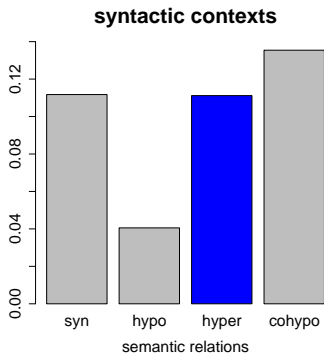
4. Results: quality of the models



4. Results: quality of the models

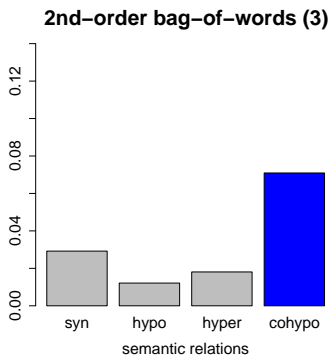
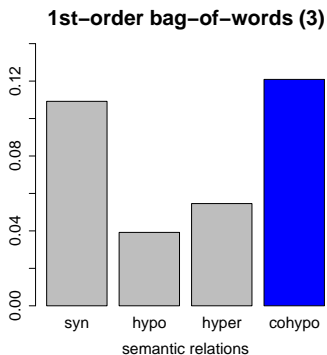


4. Results: distribution of relations



⇒ Syntactic models find more hypernyms

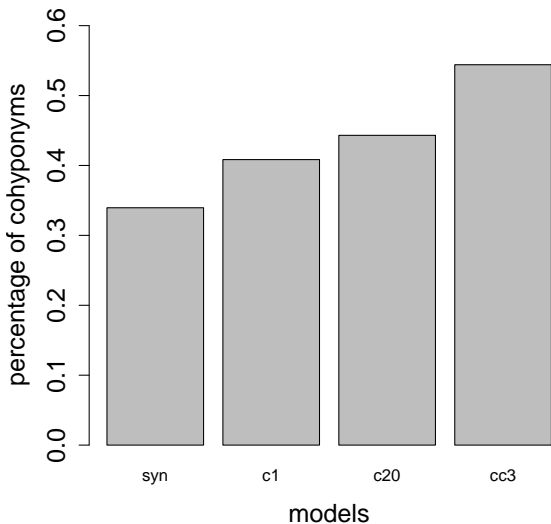
4. Results: distribution of relations



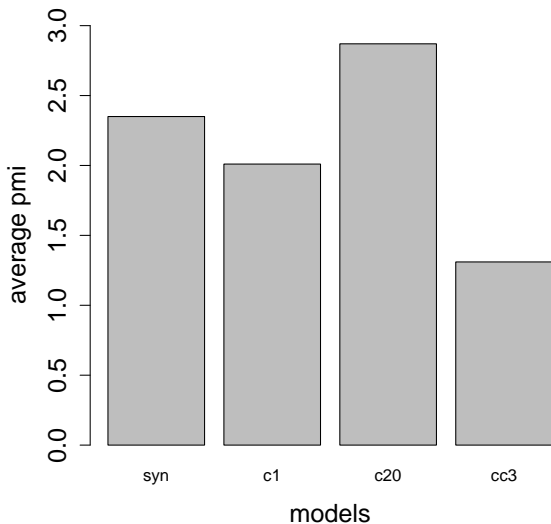
⇒ 2nd-order models have a bias towards co-hyponyms



4. Results: relative frequency of cohyponyms



4. Results: average PMI



Overzicht

1. Introduction
2. Experimental setup
3. Evaluation
4. Results
5. Conclusions



5. Conclusions

Research question

What semantic similarity is captured by distributional word spaces?

Average Wu & Palmer score

Syntactic contexts > 1st-order bag of words small contexts >
1st-order bag of words large contexts > 2nd-order bag of words

Relative frequency of cohyponyms

2nd-order bag of words > 1st-order bag of words large contexts >
1st-order bag of words small contexts > syntactic contexts





For more information:

<http://wwling.arts.kuleuven.be/qlvl>
kris.heylen@arts.kuleuven.be
yves.peirsman@arts.kuleuven.be