Size Matters

Tight and Loose Context Definitions in English Word Space Models

Yves Peirsman, Kris Heylen and Dirk Geeraerts

KULeuven
Quantitative Lexicology and Variational Linguistics

What happened before...

- Tim: syntactic models $>$ bag-of-word models
  - $\rightarrow$ What happens if we change the context size?
- John: small contexts $>$ large contexts
  - $\rightarrow$ Focus on similarity tasks.
- Marco: context size influences relationship that is modelled
  - $\rightarrow$ small contexts $\sim$ similarity
  - $\rightarrow$ large contexts $\sim$ topical relations

$\Rightarrow$ Study the influence of context size for drastically different tasks.

What happened before...

- Tim: syntactic models $>$ bag-of-word models
  - $\rightarrow$ What happens if we change the context size?
- John: small contexts $>$ large contexts
  - $\rightarrow$ Focus on similarity tasks.
- Marco: context size influences relationship that is modelled
  - $\rightarrow$ small contexts $\sim$ similarity
  - $\rightarrow$ large contexts $\sim$ topical relations

$\Rightarrow$ Study the influence of context size for drastically different tasks.

What happened before...

- Tim: syntactic models $>$ bag-of-word models
  - $\rightarrow$ What happens if we change the context size?
- John: small contexts $>$ large contexts
  - $\rightarrow$ Focus on similarity tasks.
- Marco: context size influences relationship that is modelled
  - $\rightarrow$ small contexts $\sim$ similarity
  - $\rightarrow$ large contexts $\sim$ topical relations

$\Rightarrow$ Study the influence of context size for drastically different tasks.

What happened before...

- Tim: syntactic models $>$ bag-of-word models
  - $\rightarrow$ What happens if we change the context size?
- John: small contexts $>$ large contexts
  - $\rightarrow$ Focus on similarity tasks.
- Marco: context size influences relationship that is modelled
  - $\rightarrow$ small contexts $\sim$ similarity
  - $\rightarrow$ large contexts $\sim$ topical relations

$\Rightarrow$ Study the influence of context size for drastically different tasks.

Outline

Outline

1. Introduction

Bag-of-word models

model the meaning of a word in terms of its context words in a corpus.

Context parameters

- Size of the context window: 1, 2, 3, 4, 5, 7, 10
- Order of the context words:
  - First order: what words appear in the context of the target?
  - Second order: what words appear in the context of the target's context words?

1. Introduction

Example

An accident happened just as I steered my brand new
car
on to the motorway to Hamburg yesterday.

1. Introduction

Example

An accident happened just as I steered my brand new car on to the motorway to Hamburg yesterday.

1. Introduction

Example

An accident happened just as I steered my brand new car
on to the motorway to Hamburg yesterday.

1. Introduction

Example

An accident happened just as I steered my brand new car
on to the motorway to Hamburg yesterday.

1. Introduction

Example

	truck		accident			car		
	victim		see			wheel		
	cause		yesterday			road		
An	accident	happened	just	as	I	steered	my	brand new

car

on to the	motorway to	Hamburg	yesterday.
	car	Berlin	tomorrow
	truck	live	work
	drive	harbour	remember

1. Introduction

Hypothesis

- Tighter (small and first-order) contexts find semantically similar words
  - → car–truck, sparrow–pigeon, book–novel,...
- Looser (larger and second-order?) contexts are biased towards topically related words
  - → doctor–hospital, hand–finger, car–wheel

Tasks

- Semantic similarity: word clustering
- Semantic/topical relatedness: free association norms?

1. Introduction

Parameter Settings

- Data: BNC, 100 million words, lemmatized and PoS-tagged
- Dimensionality: 5,000
- Cut-off: increasing with context size
- Stoplist: yes
- Weighting scheme: point-wise mutual information
- Similarity measure: cosine

Outline

2. Semantic similarity

Word clustering tasks

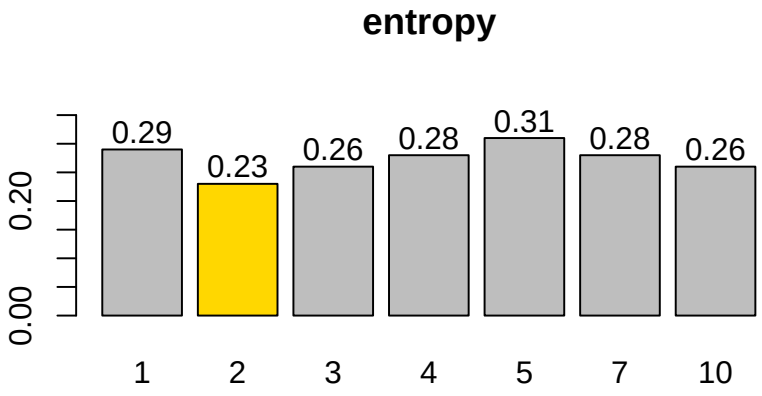
- concrete nouns: *tools*, *fruit*, *birds*, etc.
- concrete vs abstract nouns
- verbs: *communication*, *mental state*, etc.

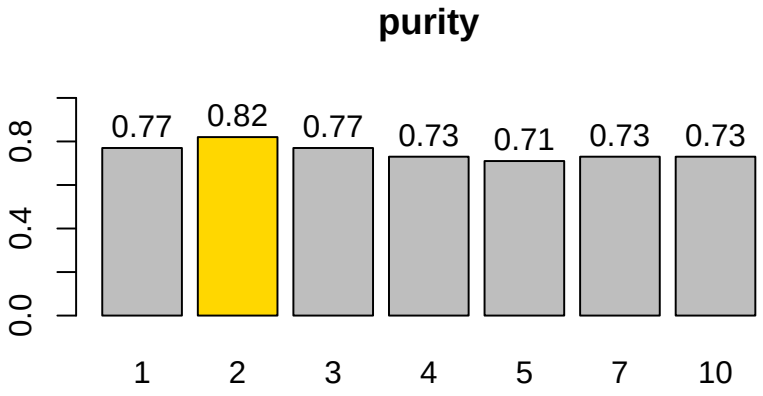
Evaluation

- Entropy: the "uncertainty" of the clustering solution
- Purity: the average proportion of a cluster taken up by the largest class
- Focus on first-order models

2. Semantic similarity

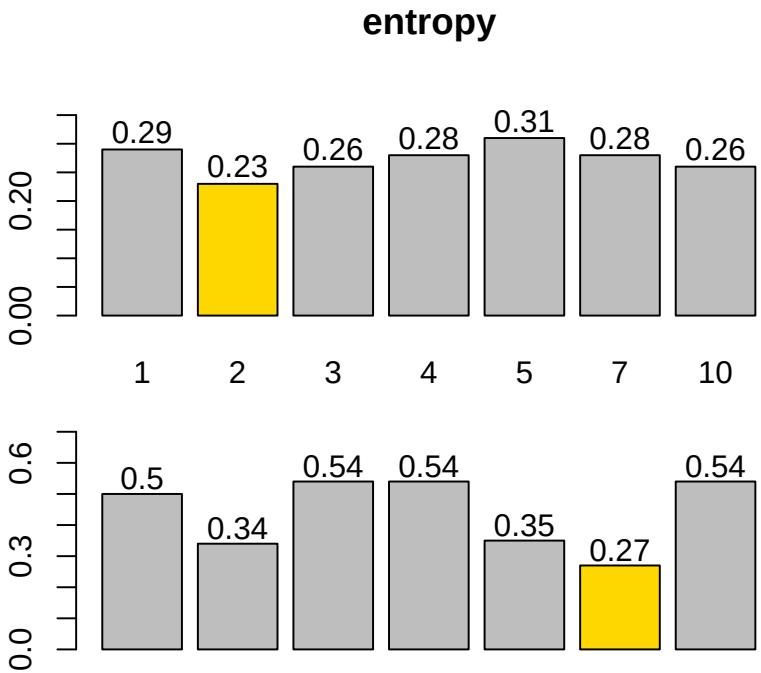
Task 1a: Concrete nouns

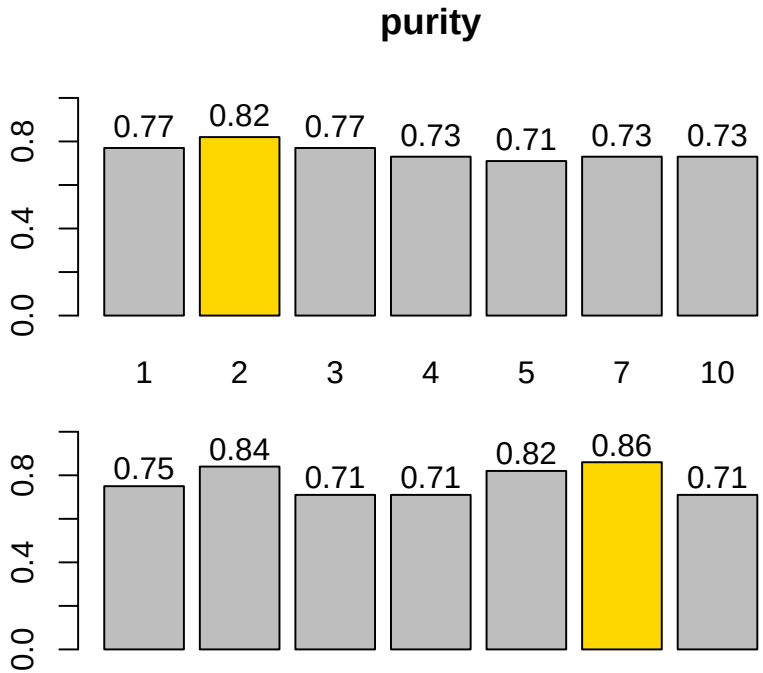




2. Semantic similarity

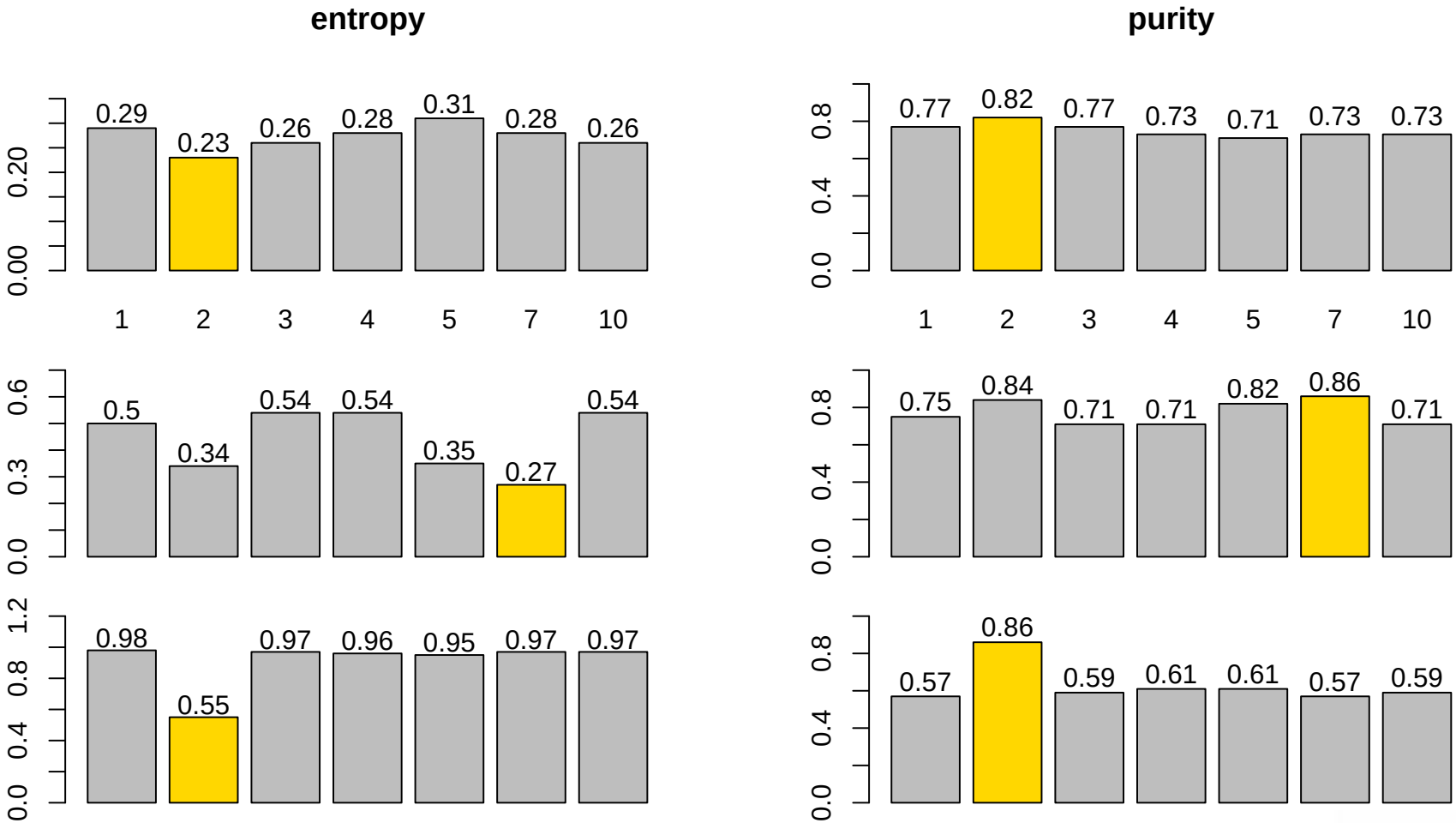
Task 1a: Concrete nouns





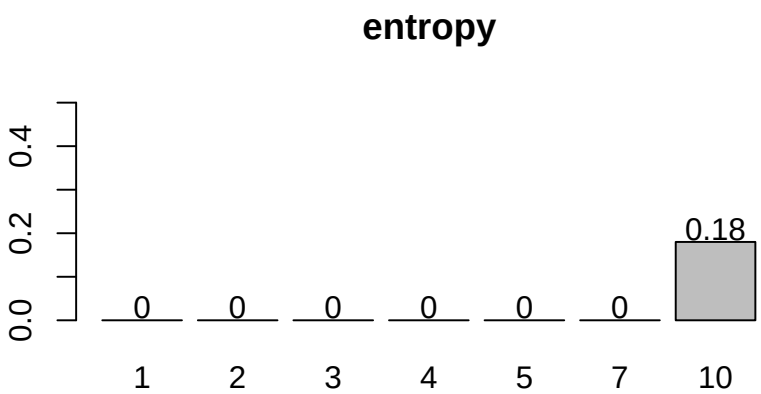
2. Semantic similarity

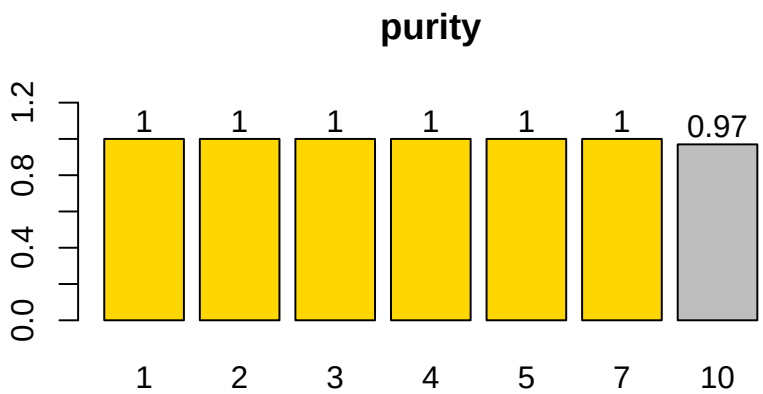
Task 1a: Concrete nouns



2. Semantic similarity

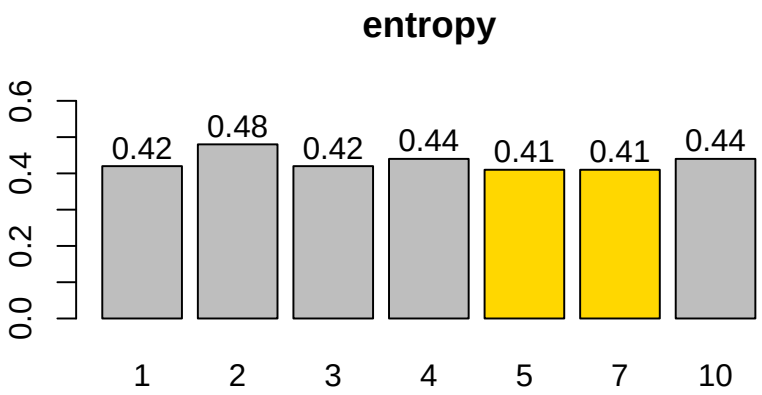
Task 1b: Concrete and abstract nouns

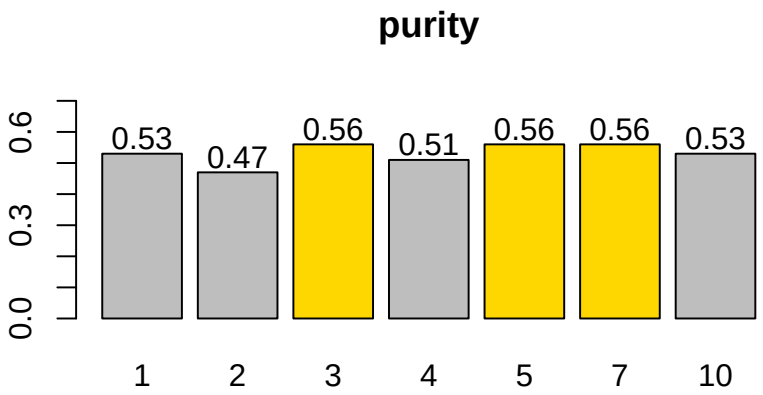




2. Semantic similarity

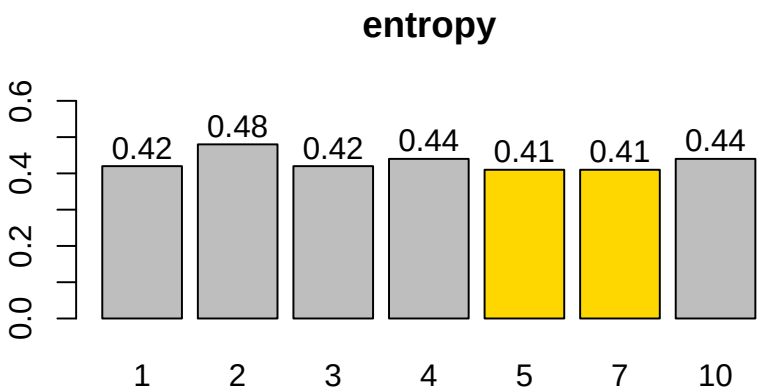
Task 1c: Verbs

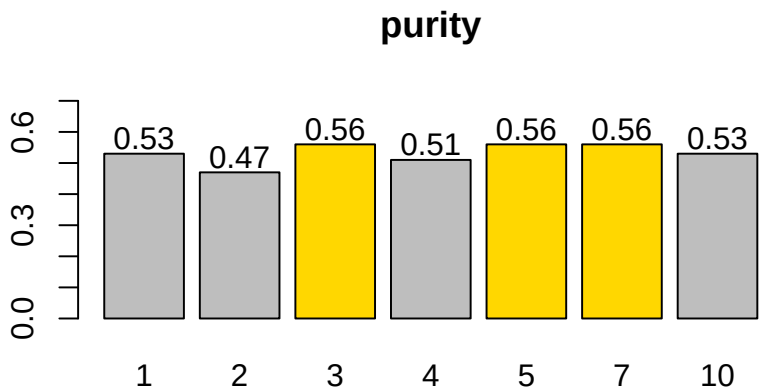


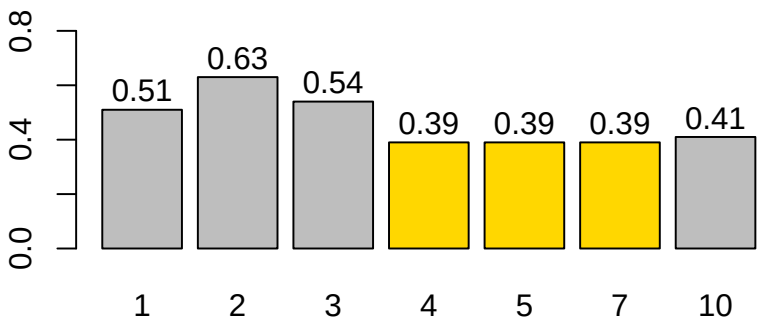


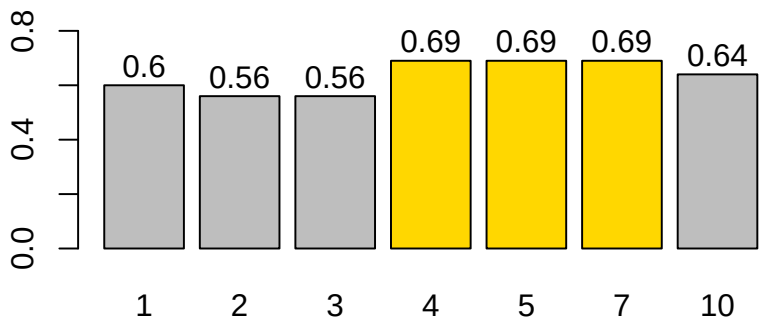
2. Semantic similarity

Task 1c: Verbs









2. Semantic similarity

Evaluation: nouns

- Overall smaller context sizes score best.
- Problematic categories: fruit vs vegetables, ground animals vs birds, tools
- Often a "kitchen" cluster emerges.

Evaluation: verbs

- Overall intermediate context sizes score best.
- Differences between (fuzzy) classes often very subtle.
  - e.g., change location (move) vs motion manner (run)

Outline

3. Associations

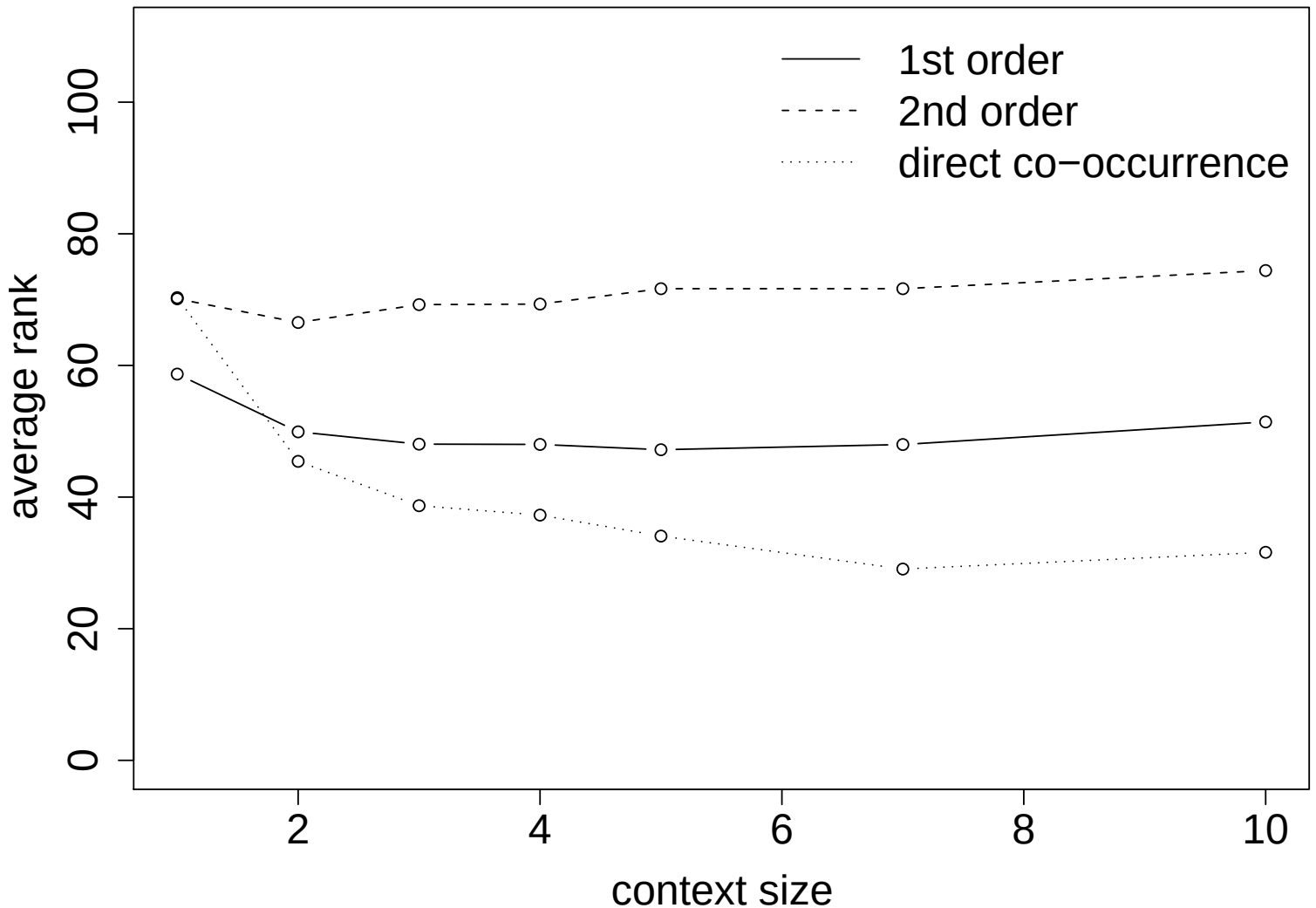
Association norms

- Each word is semantically associated with many other words.
  - e.g., pepper–salt, wave–sea, twentieth–century
- Associations are a mixture of paradigmatically and syntagmatically related words.

Task

- For each word in the test set, find the most frequent association.
- Candidates: 10,000 most frequent words in the BNC.
- The lower the average rank of the association in the 100 most related words, the better.

3. Associations



3. Associations

Quantitative evaluation

- First-order models perform much better than second-order models.
- Intermediate context sizes give the best results.
- Direct co-occurrence statistics (e.g., log-likelihood) clearly outperform word space models!

Qualitative evaluation

- Quantitative differences within first-order models are small.
- Type of recovered associations depends on context size.
- Biggest differences in ranks between context sizes 1 and 10.

3. Associations

Largest positive difference in ranks for context size 10

cue	asso	diff	cue	asso	diff
sill	window	100	damsel	distress	97
riding	horse	100	leash	dog	96
reflection	mirror	100	consultant	doctor	95
nigger	black	100	pram	baby	94
hoof	horse	100	barrel	beer	94
holster	gun	100	twentieth	century	91
dump	rubbish	100	handler	dog	90
spend	money	98	scissors	cut	80
bidder	auction	98	deck	ship	75
wave	sea	97	suicide	death	72

3. Associations

Largest positive difference in ranks for context size 1

cue	asso	diff	cue	asso	diff
melancholy	sad	100	glucose	sugar	63
rapidly	quickly	98	fund	money	61
plasma	blood	95	suspend	hang	61
astonishment	surprise	91	adequate	enough	54
joyful	happy	83	levi	jeans	49
hard	soft	78	sugar	sweet	46
cormorant	bird	76	din	noise	44
new	old	70	no	yes	42
combat	fight	69	tumour	brain	39
wrath	anger	64	weary	tired	33

3. Associations

Discussion

- Small and larger context sizes have different preferences:
  - paradigmatic similarity for context size 1
  - syntagmatic relatedness for context size 10
- Intermediate context sizes perform best, probably because they strike a balance.
- Direct co-occurrence statistics make word space models unnecessary here.

Outline

4. Conclusions

Influence of context size on discovery of semantic relations

- Word clustering task
  - nouns: small context sizes (2)
  - verbs: larger context sizes (4–7)

  $\Rightarrow$ Success of smallest context sizes mainly true for nouns
- Association norms
  - Best results for log-likelihood statistic
  - Clear difference in preferences between small and larger context sizes
  - Association norms contain many different relations

  $\Rightarrow$ combinations of different context sizes? committees?
- Student session presentation next Thursday

For more information:

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

yves.peirsman@arts.kuleuven.be