Syntactic Contexts for Finding Semantically Similar Words

CLIN
December 17th 2004

Gosse Bouma
Lonneke van der Plas
Humanities Computing
University of Groningen

Motivation(1)

Overview

Motivation

Distributional similarity

Similarity Measures

Data

Evaluation

Results

Conclusion

- Extending Dutch EuroWordNet for:
  - Question Classification

    `Welke tennisser ...?` > person ques.

  - Answering 'which' questions

    `Welk beroep heeft Renzo Piano?`

    De **Italiaan** Renzo Piano is **architect.**

Motivation(2)

- Disambiguation
  - Coordination:

(Chirac uit Frankrijk) en Blair.

Chirac uit (Frankrijk en Blair).

  - PP-attachment

Hij at (mie) met stokjes.

Hij at (mie met stokjes).

Overview

Motivation

Distributional similarity

Similarity Measures

Data

Evaluation

Results

Conclusion

Overview

Motivation

Distributional similarity

Similarity Measures

Data

Evaluation

Results

Conclusion

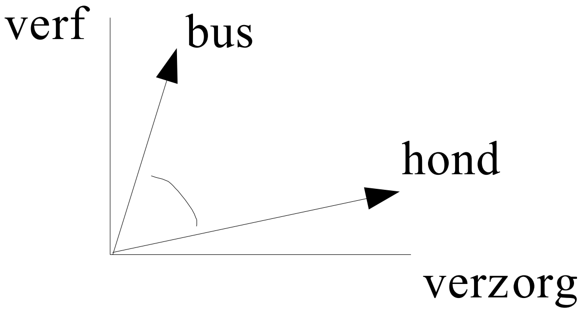
Distributional Similarity

- Similar words share similar contexts
- Context = Syntactic context

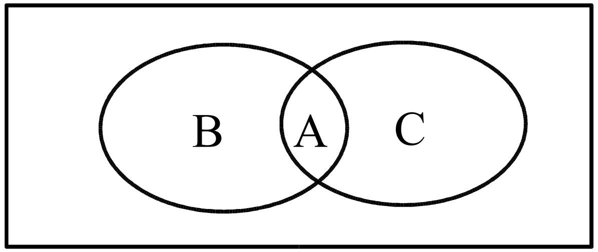
	zie	verf	verzorg	laat_uit
bus	50	5	1	0
hond	56	1	5	8

Similarity Measures

- Cosine



- Dice



$$\frac{2A}{2A+B+C}$$

Overview

Motivation

Distributional similarity

Similarity Measures

Data

Evaluation

Results

Conclusion

Weights

- Mutual Information
- T-Test

	hebben	verdoen	geven	doordrijven
zin	500	0	400	18
tijd	560	10	600	0

Overview

Motivation

Distributional similarity

Similarity Measures

Data

Evaluation

Results

Conclusion

Data

- 78 million words of parsed Dutch newspaper text

Subject-Verb	kat eet
Verb-Object	voer kat
Adjective-Noun	langharige kat
Coordination	Bassie en Adriaan
Apposition	de clown Bassie
PC	begin_met werk

Overview

Motivation

Distributional similarity

Similarity Measures

Data

Evaluation

Results

Conclusion

Extracted from data

Gram rel.	# types	# tuples
Coord	2.465.098	965.296
Subj	2.122.107	5.639.140
Adj	1.040.785	3.262.403
Obj	993.913	2.642.356
Appo	526.337	602.970
PC	389.139	770.631

Tuples,triples etc

	ga_subj	geel_adj	neem_obj	Lassie_app
bus	4	9	8	0
hond	4	1	6	8

Cutoff:
row > 10

Example Output

(for combi obj/adj using MI and Dice)

Trainingspak

bloes

tricot

tshirt

hemd

shirt

tenue

winterjas

bretel

Overview

Motivation

Distributional similarity

Similarity Measures

Data

Evaluation

Results

Conclusion

Example Output

(for subject using TTest and Cosine)

Aanduiding

```
quorum
skiluik
term
dashboardkastje
raambediening
Pep
godfather
gelijkwaardigheid
```

Example Output

(for coord. rel. using MI and Dice)

Wim Kok

Elco Brinkman

Thijs Wöltgens

Hans van Mierlo

Frits Bolkestein

Jacques Wallage

Relus ter Beek

Hedy d'Ancona

Felix Rottenberg

Overview

Motivation

Distributional similarity

Similarity Measures

Data

Evaluation

Results

Conclusion

Evaluation Framework

- Dutch EWN as gold standard
  - Advantages
    - No need for human evaluators, nor application
  - Disadvantages:
    - EWN contains no proper names
    - More than one way to measure distance in WordNet

Evaluation Framework

- 1000 words from EWN (random, freq> 10)

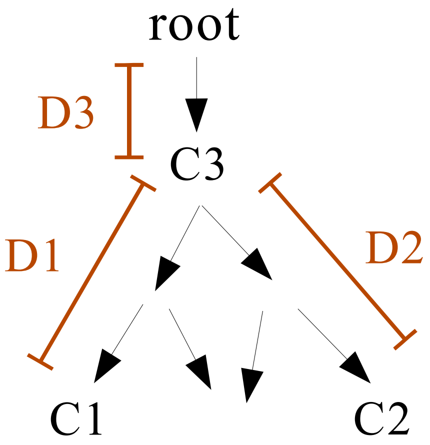
For each word

→ collect 100 most similar words

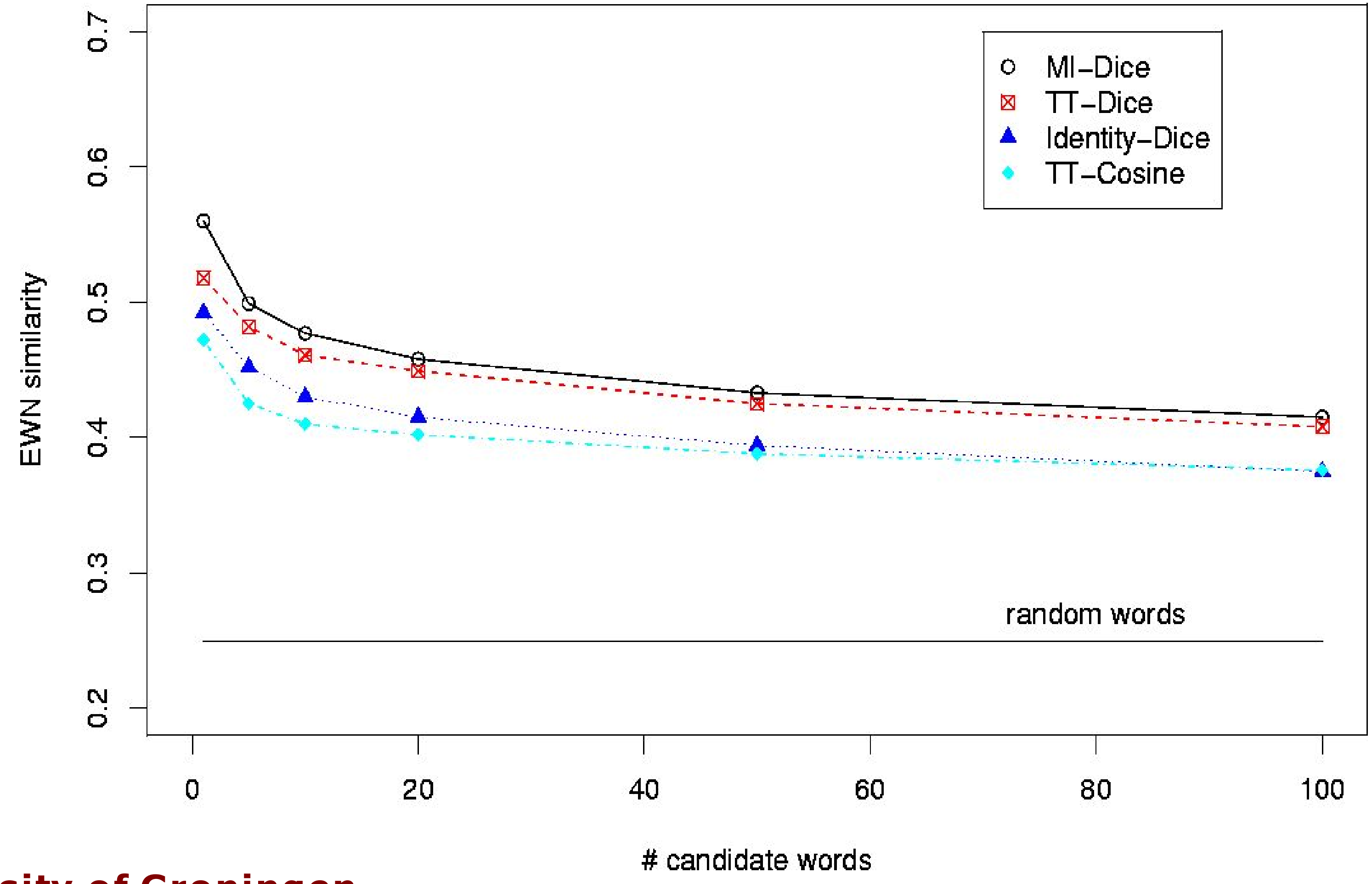
For each pair of words

→ measure similarity in EWN with Wu & Palmer's measure

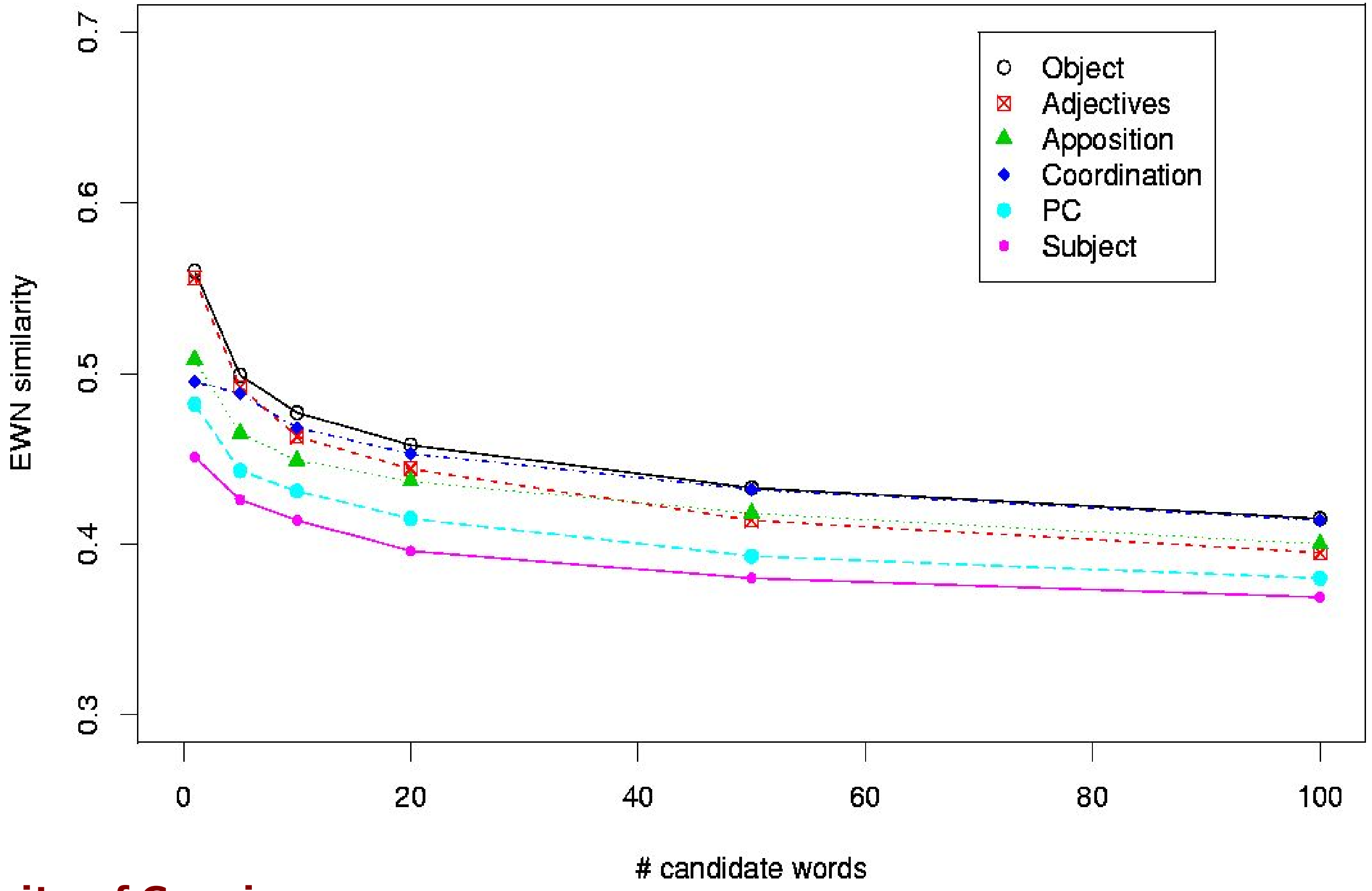
$$\frac{2(D3)}{D1+D2+2(D3)}$$



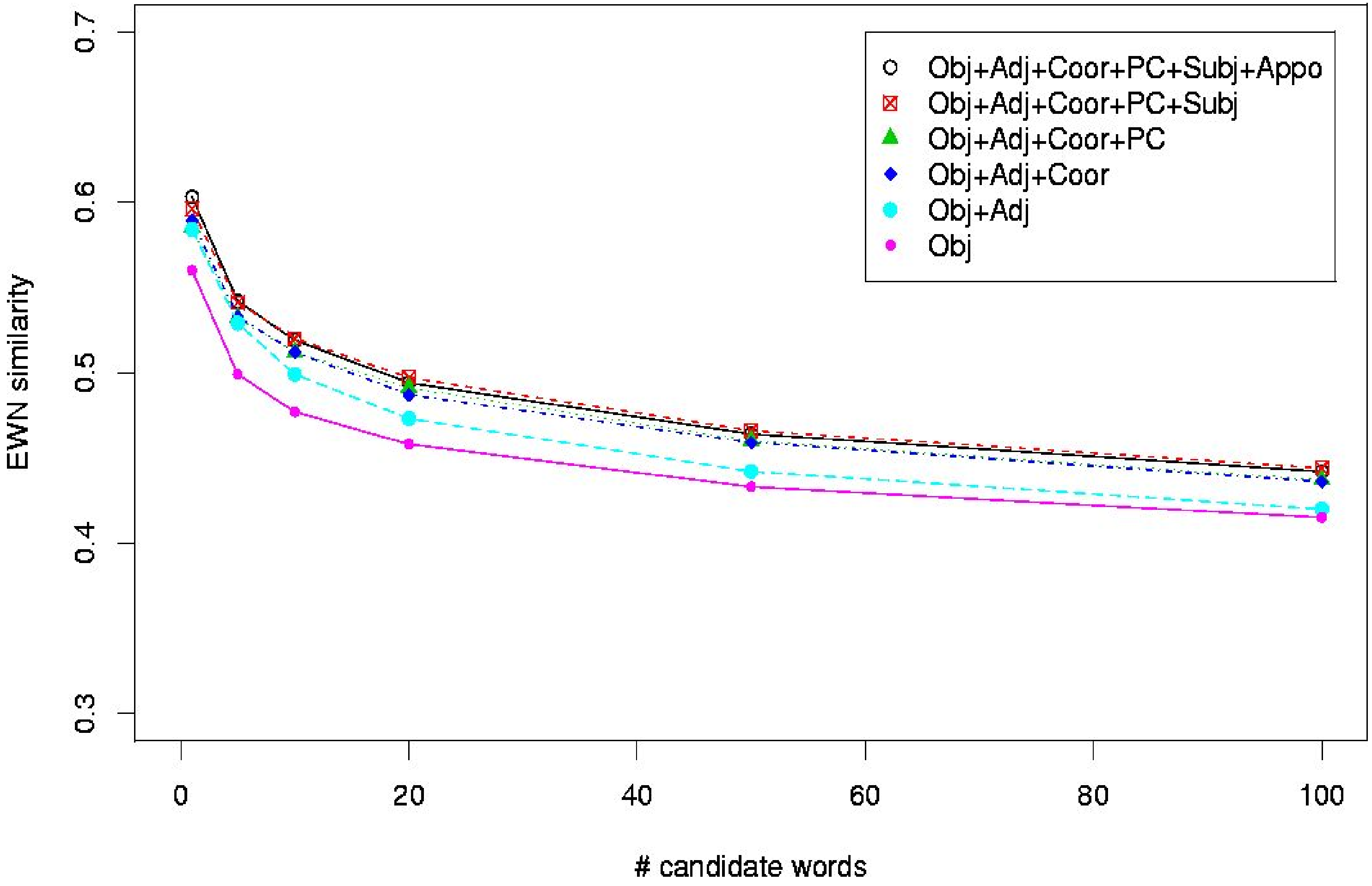
Results (object relation)



Results (diff. gram rels Dice MI)



Results (combining gramrels)



Overview

Motivation

Distributional similarity

Similarity Measures

Data

Evaluation

Results

Conclusion

Recall

Gram rel	Recall (%)
Appo	11,2
PC	29,8
Obj	47,8
Adj	50,3
Coord	56,0
Subj	57,9

Gram rel	Recall (%)
O+A	62,3
O+A+C	72,9
O+A+C+P	74,5
O+A+C+P+S	78,8
All	78,9

- Recall is number of words found in our matrix
- On average 40 % of target's most similar words are not found in EWN

Conclusions

- Combination of Dice and MI gives best results.
- Object seems most suitable grammatical relation for this task, although evaluating on EWN is not fair for Coord + Appo
- Combining grammatical relations improves results.
- With 40 % of words missing in EWN room for improvement

Overview

Motivation

Distributional similarity

Similarity Measures

Data

Evaluation

Results

Conclusion

Future Work

- From horizontal to vertical relations (hypernym structures) using
  - Clustering techniques
  - Patterns from encyclopaedia
  - Better similarity measures
- LSA, LDA to compress feature space, allow for generalization
- Get more insight into behaviour of different grammatical relations