# A quantitative evaluation of semantic word space models

Tim Van de Cruys

University of Groningen

CLIN 20
February 5, 2010
Utrecht

university of groningen

# Distributional similarity

Word space models are able to extract similar words from text:

- **Utrecht**: *Groningen*, *Amersfoort*, *Eindhoven*, *Nijmegen*, *Arnhem*, *Den Haag*, *Rotterdam*, *Zwolle*, *Amsterdam*, *Tilburg*
- **conferentie** 'conference': *bijeenkomst* 'meeting', *congres* 'conference', *symposium* 'symposium', *vergadering* 'meeting', *VN-conferentie* 'UN conference', *klimaatconferentie* 'climate conference', *topconferentie* 'summit', *forum* 'forum', *hoorzitting* 'hearing', *studiedag* 'seminar'
- **liefde** 'love': *vriendschap* 'friendship', *verlangen* 'desire', *passie* 'passion', *verdriet* 'sadness', *eenzaamheid* 'loneliness', *respect* 'respect', *angst* 'fear', *geluk* 'happiness', *jaloezie* 'jealousy', *haat* 'hate'


university of groningen

# Different kinds of context

- Three different word space models based on context:
  - document-based model (nouns $\times$ documents)
  - window-based model (nouns $\times$ context words)
  - syntax-based model (nouns $\times$ dependency relations)
- Each model with plethora of parameters!

# Different kinds of semantic similarity

- **'tight', synonym-like similarity**: (near-)synonymous or (co-)hyponymous
- **loosely related, topical similarity**: more loose relationships, such as association and meronymy

# Different kinds of semantic similarity

- **'tight', synonym-like similarity**: (near-)synonymous or (co-)hyponymous
- **loosely related, topical similarity**: more loose relationships, such as association and meronymy

## Example

- **arts** 'doctor': *dokter* 'doctor', *medicus* 'doctor', *huisarts* 'family doctor', *chirurg* 'surgeon', *specialist* 'specialist', *gynaecoloog* 'gynaecologist'
- **arts** 'doctor': *patiënt* 'patient', *ziekte* 'disease', *diagnose* 'diagnosis, *behandeling* 'treatment, *ziekenhuis* 'hospital', *stethoscoop* 'stethoscope'

# Two research questions

1. What kind of semantic similarity is captured by different context models?
2. Which models perform best?
   - context
   - document size, context window size
   - weighting function
     - logarithmic
     - entropy
     - pointwise mutual information
   - $\pm$ dimensionality reduction
     - singular value decomposition
     - non-negative matrix factorization

university of groningen

# Evaluation framework

- Compare results to Dutch CORNETTO database
- Two similarity measures:
  - path length: Wu & Palmer similarity measure
  - information theoretic: Lin's similarity measure
- Nouns close in the hierarchy are tightly similar
- Pairwise similarity for $k$ similar words
- Test set of $\pm$ 5000 nouns

# Results 1/2

| model | wu & palmer's similarity $k=1$ | $k=3$ | $k=5$ | lin's similarity $k=1$ | $k=3$ | $k=5$ |
|---|---|---|---|---|---|---|
| document | .379 | .331 | .309 | .354 | .320 | .304 |
| window$_{(w=par)}$ | .442 | .377 | .349 | .404 | .357 | .336 |
| window$_{(w=2)}$ | .633 | .561 | .526 | .541 | .485 | .456 |
| syntax | **.648** | **.584** | **.554** | **.555** | **.504** | **.480** |
| baseline | .128 | .126 | .126 | .164 | .163 | .163 |

$$\text{syntax} > \text{window}_{(w=2)} \gg \text{window}_{(w=par)} > \text{document}$$

university of groningen

# Results 2/2

- Syntax (with PMI) best
- Closely followed by small window (with PMI)
- Large window and document perform much worse
- dimensionality reduction only helps to improve document-based (a little)

# Evaluation framework

- Compare output of clustering algorithm to gold standard classification
- Two cluster tasks (ESSLLI 2008 workshop's shared task)
  - concrete noun categorization (44 nouns)
    - 2-way: *natural*, *artefact*
    - 3-way: *animal*, *vegetable*, *artefact*
    - 6-way: *bird*, *groundAnimal*, *fruitTree*, *green*, *tool*, *vehicle*
  - abstract/concrete noun discrimination (30 nouns)
    - 2-way: HI, LO
- Evaluation measures:
  - Entropy: distribution of classes within cluster (small = good)
  - Purity: ratio of largest class present in cluster (large = good)

university of groningen

# Results 1/2

| | 2-way | | 3-way | | 6-way | |
|---|---|---|---|---|---|---|
| model | ent | pur | ent | pur | ent | pur |
| document | .930 | .614 | .179 | .932 | .292 | .682 |
| window$_{(w=par)}$ | .911 | .659 | .541 | .705 | .377 | .591 |
| window$_{(w=2)}$ | **.000** | **1.000** | .213 | .909 | .206 | .773 |
| syntax-based | **.000** | **1.000** | **.000** | **1.000** | **.153** | **.864** |

$$\text{syntax} > \text{window}_{(w=2)} \gg \text{window}_{(w=par)} > \text{document}$$

university of groningen

# Results 2/2

- Same tendencies as wordnet-based similarity
- model with large window seems to extract topically related clusters:
  - *aardappel* 'potato', *ananas* 'pineapple', *banaan* 'banana', *champignon* 'mushroom', *fles* 'bottle', *kers* 'cherry', *ketel* 'kettle', *kip* 'chicken', *kom* 'bowl', *lepel* 'spoon', *peer* 'pear', *sla* 'lettuce', *ui* 'onion'
- Similar result for abstract/concrete noun discrimination

university of groningen

# Evaluation framework

- Coherence of semantic domain tags (available in CORNETTO)
- 'particular areas of human knowledge' (POLITICS, MEDICINE, SPORTS)
- $\longrightarrow$ topical similarity
- Ratio of most frequent domain tag (also in tagset of target word) over top 10 similar words
- Same test set of $\pm$ 5000 nouns

# Results 1/2

| model | $sim_{topic}$ |
|---|---|
| document | .394 |
| window$_{(w=par)}$ | .399 |
| window$_{(w=2)}$ | .414 |
| syntax | **.441** |
| baseline | .048 |

$$\text{syntax} > \text{window}_{(w=2)} \cong \text{window}_{(w=par)} \cong \text{document}$$

university of groningen

# Results 2/2

- Syntax still scores best
- Other models do not perform much worse
- No real difference between small window and large window
- $\longrightarrow$ large window and document do not extract tight similarity, but they do grasp topical similarity

university of groningen

# Conclusion

- Different context leads to different kind of similarity
- Syntax, small window $\leftrightarrow$ large window, documents
- Former models perform well on both kinds of similarity, latter models only extract topically similar words