Classification of Noun-Noun Compound Semantics in Dutch and Afrikaans

Ben Verhoeven¹, Walter Daelemans¹ & Gerhard van Huyssteen²

¹ CLiPS – CLG, University of Antwerp, Belgium
{Ben.Verhoeven;Walter.Daelemans}@ua.ac.be

² CTexT, NWU, Potchefstroom, South Africa
Gerhard.VanHuyssteen@nwu.ac.za

Presented at CLIN 2013
Enschede, The Netherlands

Universiteit Antwerpen
Introduction

• Productivity of a language to create new words
  - Obstacle for computational language understanding
• Meaning of compound is often not clear on its own (ambiguity)
• Implicit semantic relation between constituents
  - e.g. donut seat
    • ‘donut-shaped seat’
    • ‘seat with a donut nearby’
    • ‘seat made of donuts’ ?
Applications

- Natural language understanding
  - Machine translation
    - Paraphrase may be needed
    - e.g. *Antwerp hostel* (Eng) -> *Auberge à Anvers* (Fr)
  - Information retrieval
  - Information extraction
  - Question answering
Related Research (1)

• Focus on
  - English
  - Noun-noun compounds

• Supervised machine learning problem
• Predefined inventory of classes of semantic relations between constituents of compound
Related Research (2)
Classification

• Two kinds of classification schemes
  - Paraphrasing preposition
    • E.g. *autodeur* = deur VAN auto
  - Predicate-based classes
    • Class AGENT: ‘X is performed by Y’
      - E.g. *studentenprotest* = protest performed by students
Related Research (4)

Features

• Taxonomy-based methods
  - Semantic network similarity
  - Word’s location in hierarchy of terms
    • E.g. Hyponomy in WordNet
      - E.g. cola < frisdrank < drank < vloeistof

• Corpus-based methods
Related Research (5) Features

• Taxonomy-based methods
• Corpus-based methods
  - Co-occurrence information of constituents in corpus
  - Distributional hypothesis (Harris)
    • Set of contexts in which a word occurs is an implicit representation of its semantics
Annotation (1)

- Semantic information on compounds needed for machine learning
- Explicit description by manual annotation
- Constraints on compound selection
  - Not in dictionary
    - Otherwise, gloss already present
    - Train classifier on systematics of newly produced compounds
  - Constituents in dictionary
    - Semantically relating of unknown words seems pointless
Annotation (2)
Scheme and Guidelines

• Adopted from Ó Séaghdha (2008), adapted for Afrikaans and Dutch
• 11 classes of compounds that describe relation between constituents
• Of which 6 semantically specific
  - BE e.g. zanger-muzikant skrywer-boer
  - HAVE autodeur blomsteel
  - IN tuinfeest nagaktiwiteite
  - ACTOR studentenprotest beerjagter
  - INST hamerslag tapytborsel
  - ABOUT postzegelverzameling kategismusvrae
Dutch
• Compound list from e-Lex
• 1802 noun-noun compounds
• Second annotator: 500
• IAA = 60.2 %
  (Kappa = 0.60)

Afrikaans
• 1500 noun-noun compounds manually selected from Ckarma
• 3 annotators
• IAA = 53.4%
  (Kappa = 0.53)
Experiment (1)

- Ó Séaghdha (2008) as inspiration

- Lexical similarity
  - Compounds are semantically similar when their respective constituents are semantically similar
  - E.g. *mieliesak* ‘corn bag’ and *graanblik* ‘can of grain’
Experiment (2) Vector Creation

• Co-occurrence context for every compound constituent
  - For each instance of constituent, \( n \) surrounding words were held in memory
  - Size of context: 3 & 5 left and right (Dutch also 1, 2 & 4)
  - Relative frequencies of context words stored in vector

• Twente News Corpus (Dutch): 340 million words
• Taalkommissiekorpus (Afrikaans): 60 million words

Universiteit Antwerpen
Experiment (3)  
Vector Creation

- Instance vectors are concatenation of constituent data
- Relative frequencies for the 1000 most frequent words per constituent (2000 per compound)
- Experiment only on compounds in semantically specific classes
  - BE, HAVE, ABOUT, IN, ACTOR, INST
Principal Component Analysis (PCA)

- Size of vectors: 2000 attributes
- Computationally expensive
- PCA mathematically reduces dimensionality while optimising variance in data
- Correlated attributes are fused into principal components (PCs)
- For now: restriction to 50 PCs
Baseline

- First research for these languages
- Majority baseline, thus:
  - For Dutch: 29.5% (428/1447 class IN)
  - For Afrikaans: 28.2% (407/1439 class ABOUT)
Initial Results

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOW 3</td>
<td>47.1</td>
<td>47.9</td>
<td>47.3</td>
</tr>
<tr>
<td>BOW 5</td>
<td>46.7</td>
<td>47.8</td>
<td>47.1</td>
</tr>
<tr>
<td>PCA 3</td>
<td>43.7</td>
<td>47.3</td>
<td>43.7</td>
</tr>
<tr>
<td>PCA 5</td>
<td>42.9</td>
<td>48.0</td>
<td>43.2</td>
</tr>
<tr>
<td>Baseline</td>
<td>29.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOW 3</td>
<td>50.8</td>
<td>51.6</td>
<td>51.1</td>
</tr>
<tr>
<td>BOW 5</td>
<td>50.3</td>
<td>50.8</td>
<td>50.5</td>
</tr>
<tr>
<td>PCA 5</td>
<td>49.3</td>
<td>51.3</td>
<td>48.5</td>
</tr>
<tr>
<td>PCA 3</td>
<td>47.7</td>
<td>50.5</td>
<td>47.5</td>
</tr>
<tr>
<td>Baseline</td>
<td>28.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Results of SVM on Dutch and Afrikaans compound semantics, using 10-fold cross-validation
- BOW and PCA[50]
- Size of context: 3 & 5

Universiteit Antwerpen
Initial Discussion

• Both languages show significant improvement over majority baseline
• BOW seems to do better than PCA

• Better results for Afrikaans
  - Possibly due to annotated list being a combination of semantic annotations of 3 persons
  - Most agreed upon class for each compound
• Dutch: just one annotator
More experiments for Dutch

- Selection of context words considered
  - All words (BOW)
  - Only content words (verbs, nouns, adjectives and adverbs) (VNA)
  - Only function words (determiners, prepositions, conjugations, pronouns) (Func)

- PCA: calculation of more PCs
<table>
<thead>
<tr>
<th>AVG</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOW</td>
<td>46.50</td>
</tr>
<tr>
<td>VNA</td>
<td>46.24</td>
</tr>
<tr>
<td>Func</td>
<td>45.70</td>
</tr>
<tr>
<td>1</td>
<td>44.58</td>
</tr>
<tr>
<td>2</td>
<td>45.57</td>
</tr>
<tr>
<td>3</td>
<td>45.87</td>
</tr>
<tr>
<td>4</td>
<td>45.72</td>
</tr>
<tr>
<td>5</td>
<td>45.87</td>
</tr>
<tr>
<td>PCA[50]</td>
<td>43.64</td>
</tr>
<tr>
<td>PCA[100]</td>
<td>45.18</td>
</tr>
<tr>
<td>PCA[150]</td>
<td>45.86</td>
</tr>
<tr>
<td>Baseline</td>
<td>29.50</td>
</tr>
</tbody>
</table>
Discussion

• Hardly any difference using VNA or Func
• BOW maintains best results

But:
• PCA using 150 PCs approaches BOW results
  - Significant improvement over 50 PCs

• Context size:
  - 1 seems not enough
  - No real differences among the rest
Per-class performance
Dutch BOW 3

<table>
<thead>
<tr>
<th>Category</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>IN</td>
<td>60.1</td>
</tr>
<tr>
<td>ABOUT</td>
<td>52.9</td>
</tr>
<tr>
<td>HAVE</td>
<td>36.3</td>
</tr>
<tr>
<td>INST</td>
<td>40.6</td>
</tr>
<tr>
<td>BE</td>
<td>17.0</td>
</tr>
<tr>
<td>ACTOR</td>
<td>42.9</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>47.3</strong></td>
</tr>
</tbody>
</table>

IN is best performing category
BE does significantly worse than others
Per-class performance

### Dutch BOW 3

<table>
<thead>
<tr>
<th>Category</th>
<th>F-Score</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>IN</td>
<td>60.1</td>
<td>29.5 %</td>
</tr>
<tr>
<td>ABOUT</td>
<td>52.9</td>
<td>26.6 %</td>
</tr>
<tr>
<td>HAVE</td>
<td>36.3</td>
<td>16.1 %</td>
</tr>
<tr>
<td>INST</td>
<td>40.6</td>
<td>16.2 %</td>
</tr>
<tr>
<td>BE</td>
<td>17.0</td>
<td>7.3 %</td>
</tr>
<tr>
<td>ACTOR</td>
<td>42.9</td>
<td>4.3 %</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>47.3</strong></td>
<td></td>
</tr>
</tbody>
</table>

### Afrikaans BOW 3

<table>
<thead>
<tr>
<th>Category</th>
<th>F-Score</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>IN</td>
<td>51.8</td>
<td>20.8 %</td>
</tr>
<tr>
<td>ABOUT</td>
<td>61.3</td>
<td>28.2 %</td>
</tr>
<tr>
<td>HAVE</td>
<td>23.9</td>
<td>9.7 %</td>
</tr>
<tr>
<td>INST</td>
<td>13.6</td>
<td>7.5 %</td>
</tr>
<tr>
<td>BE</td>
<td>56.9</td>
<td>25.0 %</td>
</tr>
<tr>
<td>ACTOR</td>
<td>62.2</td>
<td>8.8 %</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>51.1</strong></td>
<td></td>
</tr>
</tbody>
</table>

Classes with fewer instances seem harder to learn
Easily learnable class: ACTOR

Universiteit Antwerpen
Influence of constituent

Dutch PCA[150] 3

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Const 1</td>
<td>40.9</td>
<td>46.3</td>
<td>41.6</td>
</tr>
<tr>
<td>Const 2</td>
<td>39.3</td>
<td>42.7</td>
<td>38.7</td>
</tr>
<tr>
<td>Compound</td>
<td>45.2</td>
<td>48.4</td>
<td>45.6</td>
</tr>
<tr>
<td>Baseline</td>
<td></td>
<td>29.5</td>
<td></td>
</tr>
</tbody>
</table>

- Larger influence of first constituent on the semantics of the compound (modifier)
- Similar to findings in psycholinguistics where first constituent has more influence on the selection of the linking element (Krott, Schreuder & Baayen, 2002)

Universiteit Antwerpen
Learning curves (1)
Dutch BOW 3

- Seems to quickly reach a ceiling
- Better than baseline
Learning curves (2)
Afrikaans BOW 3

- Seems somewhat more promising
- Yet, curve already starts high
- Either more systematicity in annotation
- Or slightly better corpus for this purpose
Discussion

• Is accuracy of 50% relevant?
  - Compare with human judgement: IAA of 50-60%.
  - Not all mistakes are stupid
    • Sometimes incorrect annotation and correct classification
      - E.g. *parochiestelsel* ‘parish system’
        » Annotation: IN
        » Classification: ABOUT
    • Sometimes both annotation and classification are correct
      - E.g. *badkuur* ‘bath treatment’
        » Annotation: IN
        » Classification: INST
Conclusion

- Promising initial results for both languages
- Highest F-scores
  - Afrikaans 51.1% (vs. 28.2%)
  - Dutch 47.3% (vs. 29.5%)
- Indication: Compares favourably with English research with similar methods
  - Ó Séaghdha 58.8%

- More influence of modifier (first constituent) than head
- Learning curve shows need for more semantic information of compounds
Further Research

- Attempt to improve IAA by providing sample sentences during annotation and better educating the annotators (ongoing)
- Investigate taxonomy-based methods
  - Use Cornetto for Dutch
  - Afrikaans also has a small-scale WordNet
- XN compound semantics (ongoing)
Acknowledgement

Research sponsored by:
- Nederlandse Taalunie (Dutch Language Union)
- Departement of Arts and Culture (DAC) of South Africa
- National Research Foundation (NRF) of South Africa
Thank you!

For suggestions and/or questions:

Ben Verhoeven
CLiPS – Computational Linguistics Group
University of Antwerp

ben.verhoeven@ua.ac.be
http://www.clips.ua.ac.be/~ben