Clustering Semantically Similar words

PART I: Semantically Similar words

PART II: Clustering
Motivation (from QA)

Question Classification

Welke tennisser ...? > person ques.

Answering 'which' questions

Welk beroep heeft Renzo Piano?

De Italiaan Renzo Piano is architect.
Clustering Semantically Similar Words

What do we need?

PART I: Similar words

Motivation

Distributional similarity

Data

Similarity Measures

Output

Evaluation + perform.

Conclusion

PART II: Clustering

profession

architect

tennisplayer
Motivation (from parsing)

- **Disambiguation**

- **Coordination:**
  
  (Chirac uit Frankrijk) en Blair.

  Chirac uit (Frankrijk en Blair).

- **PP-attachment**

  Hij at (mie) met stokjes.

  Hij at (miek met stokjes).
Clustering Semantically Similar Words

What do we need?

PART I: Similar words
Motivation
Distributional similarity
Data
Similarity Measures
Output
Evaluation + perform.
Conclusion

PART II Clustering
Dutch EuroWordNet

- There is a resource for Dutch
- However, its coverage is not sufficient for our task.
- From 72 function questions, 21 misclassified, because of missing functions in EWN
Clustering Semantically Similar Words

Distributional Similarity

- Semantically similar words share similar contexts
- Context = Syntactic context

<table>
<thead>
<tr>
<th></th>
<th>zie</th>
<th>verf</th>
<th>verzorg</th>
<th>laat_uit</th>
</tr>
</thead>
<tbody>
<tr>
<td>bus</td>
<td>50</td>
<td>5</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>hond</td>
<td>56</td>
<td>1</td>
<td>5</td>
<td>8</td>
</tr>
</tbody>
</table>

PART I: Similar words
Motivation
Distributional similarity
Data
Similarity Measures
Output
Evaluation + perform.
Conclusion

PART II Clustering
Data

- 78 million words of parsed Dutch newspaper text

**PART I : Similar words**

Motivation

Distributional similarity

Data

Similarity Measures

Output

Evaluation + perform.

Conclusion

**PART II Clustering**
PART I: Similar words

Motivation

Distributional similarity

Data

Similarity Measures

Output

Evaluation + perform.

Conclusion

PART II Clustering

---

Extracted from data

<table>
<thead>
<tr>
<th>Gram rel.</th>
<th># tuples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subj</td>
<td>5,639,140</td>
</tr>
<tr>
<td>Adj</td>
<td>3,262,403</td>
</tr>
<tr>
<td>Obj</td>
<td>2,642,356</td>
</tr>
<tr>
<td>Coord</td>
<td>965,296</td>
</tr>
<tr>
<td>PC</td>
<td>770,631</td>
</tr>
<tr>
<td>Appo</td>
<td>602,970</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>ga_subj</th>
<th>geel_adj</th>
<th>neem_obj</th>
<th>Lassie_app</th>
</tr>
</thead>
<tbody>
<tr>
<td>bus</td>
<td>4</td>
<td>9</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>hond</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>8</td>
</tr>
</tbody>
</table>

Cutoff: row > 10
Clustering Semantically Similar Words

## Similarity Measures

- **Cosine**

![Cosine Diagram](image)

- **Dice**

\[
\text{Dice} = \frac{2A}{2A+B+C}
\]
Weights

- Mutual Information

<table>
<thead>
<tr>
<th></th>
<th>hebben</th>
<th>verdoen</th>
<th>geven</th>
<th>doordrijven</th>
</tr>
</thead>
<tbody>
<tr>
<td>zin</td>
<td>500</td>
<td>0</td>
<td>400</td>
<td>18</td>
</tr>
<tr>
<td>tijd</td>
<td>560</td>
<td>10</td>
<td>600</td>
<td>0</td>
</tr>
</tbody>
</table>
Example Output

danser:

muzikant,
musicus,
choreograaf,
danseres,
artiest,
renner,
zanger,
personage,
orkest, cabaretier, sporter, bezoeker, acteur,
vrijwilliger, theatermaker, technicus, clown,
toneelspeler, politiemens, actrice,
Example Output

Michael Jackson:

Prince,
Tina Turner,
Elton John,
Madonna,
Peter Gabriel,
Bruce Springsteen,
Genesis,
Rolling Stones,
Dire Straits,
David Bowie, U2, The Rolling Stones, Paul
Simon, Simple Minds, Beatles, Elvis Presley,
Sting, Bob Dylan, Bon Jovi, Neil Young,
Clustering Semantically Similar Words

Example Output

onvrede:
ongenoegen, ontevredenheid, frustratie, bezorgdheid, ergernis, ongerustheid, verontwaardiging, boosheid, irritatie, weerzin, wantrouwen, verwarring, onrust, onzekerheid, teleurstelling, afkeer, vrees, onbehagen, scepsis, twijfel,

PART I: Similar words

Motivation

Distributional similarity

Data

Similarity Measures

Output

Evaluation + perform.

Conclusion

PART II Clustering
Clustering Semantically Similar Words

Example output

blad:

tijdschrift, krant, weekblad, dagblad, maandblad, medium, pers, magazine, blaadje, bloem, televisie, stuk, tak, steen, materiaal, plant, stukje, artikel, boekje, schilderij,
Clustering Semantically Similar Words

Evaluation Framework

PART I: Similar words

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

Motivation

Distributional similarity

Data

Similarity Measures

Output

Evaluation + perform.

Conclusion

PART II Clustering

For each word

collect 100 most similar words according to system

For each pair of words

measure similarity in EWN using Wu & Palmer measure (1994)
Clustering Semantically Similar Words

Performance

PART I: Similar words
Motivation
Distributional similarity
Data
Similarity Measures
Output
Evaluation + perform.
Conclusion

PART II Clustering
Conclusions

- We now gave as output a ranked list of semantically similar words for each word

- We want to look at the groups that can be formed (semantic classes)
Conclusions (cntd.)

- We have as data:

<table>
<thead>
<tr>
<th>zie</th>
<th>verf</th>
<th>verzorg</th>
<th>laat_uil</th>
</tr>
</thead>
<tbody>
<tr>
<td>bus</td>
<td>50</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>hond</td>
<td>56</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>etc</td>
<td>etc</td>
<td>etc</td>
<td>etc</td>
</tr>
</tbody>
</table>

PART I: Similar words

Motivation

Distributional similarity

Data

Similarity Measures

Output

Evaluation + perform.

Conclusion

PART II Clustering
Clustering algorithms partition a set of objects into groups or clusters.

The criterion for clustering is similarity.

It's like putting your dirty clothes in piles of similar colours and similar washing instructions.
Motivation

- Exploratory data analysis
  Clustering people with language impairments according to the similarity in mistakes they make and see what groups emerge after clustering

- Generalization
  All kinds of animals will be in one cluster despite their differences
Hierarchical versus flat

PART I: Similar words

PART II: Clustering

Intro

hierarchical vs flat

soft vs hard

clash + solutions

flat clustering (hard)

soft clustering

more about senses
Hierarchical versus flat (cntd)

- It seems that hierarchical clustering is what we need to build a wordnet.
Will hierarchical clustering give us the hierarchy we want?

danser:
  - muzikant,
  - musicus,
  - choreograaf,
  - danseres,

artiest,
  - renner, zanger, personage, orkest, cabaretier,
  - sporter, bezoeker, acteur, vrijwilliger,
  - theattermaker, technicus, clown, toneelspeler,
  - politiemens, actrice,
Clustering Semantically Similar Words

Soft versus hard

- hard: Each object is assigned to only one cluster.

- Soft clustering allows degrees of membership and membership in multiple clusters as a degree in certainty.

- Disjunctive models: true multiple assignment (not just uncertainty)
Clustering Semantically Similar Words

Examples

PART I: Similar words

PART II: Clustering

Intro

hierarchical vs flat

soft vs hard

clash + solutions

flat clustering (hard)

soft clustering

more about senses
Soft versus hard (cntd)

- Language is not unambiguous and hard clustering does not seem a good way to deal with linguistic data.

- In my case polysemous words ('blad'). They belong to more than one cluster.
Clash

- In hierarchical clustering assignment is usually hard.

- In flat clustering: hard or soft
Possible solutions

1. first soft/disjunctive clustering
2. Give senses a unique ID, split-up features
3a. do hard/hierarchical clustering with senses included
3b. get hierarchical information from other source
Other sources for hierarchical information

- Patterns in free text such as Xs, Ys and other Zs
  Zs, such as Xs and Ys
  But can be subjective: Balkenende, Rutte and other disasters.
- Dictionaries, encyclopedia
- Dutch EWN
Flat clustering (term.)

- Cluster centre = centre of the M points in a cluster \( c = \text{centroid} \)

- Each component of the centroid vector is simply the average of the values for that component

<table>
<thead>
<tr>
<th></th>
<th>zie</th>
<th>verf</th>
<th>verzorg</th>
<th>laat_uit</th>
</tr>
</thead>
<tbody>
<tr>
<td>bus</td>
<td>50</td>
<td>5</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>hond</td>
<td>56</td>
<td>1</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>Centroid</td>
<td>53</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
Flat clustering (hard)

- Set of initial cluster centres
- Go through several iterations:
  - of assigning each object to closest cluster centre.
  - and recomputing the cluster centre
- Repeat until stable
Soft clustering

- The idea is that the observed data are generated by several causes = clusters

- Gaussian mixture model: For each element the clusters from flat clustering are still the dominant clusters. But each word also has some non-zero membership in other cluster.
More about senses

Example \textit{blad}:

\begin{itemize}
  \item \textit{blad}: publiceer
  \item \textit{plant}: sterf\_af
\end{itemize}
## More about senses

- **Soft clustering:**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Krant</td>
<td>0.9</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Blad</td>
<td>0.6</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>Plant</td>
<td>0</td>
<td>0.9</td>
<td></td>
</tr>
</tbody>
</table>
More about senses

- Disjunctive clustering

PART I: Similar words

PART II: Clustering

Intro

hierarchical vs flat

soft vs hard

clash + solutions

flat clustering (hard)

soft clustering

more about senses
More about senses

What I really want:

- publiceer
- krant
- blad1
- blad2
- plant
- sterf_af
Use bidirectionality

- Recap: Distributional similarity => Semantically similar words share similar contexts
- Our goal was clusters of nouns and verbs were our features
- Nouns can become features and verbs can become elements to be clustered.
Using this bi-directionality to get senses

- Would it be possible to first cluster the features (verbs)
- split nouns up according to number of clusters found in features (verbs) > senses
- and give each sense its accompanying subset of features.

PART I: Similar words
PART II: Clustering
Intro
hierarchical vs flat
soft vs hard
clash + solutions
flat clustering (hard)
soft clustering
more about senses
Clustering Semantically Similar Words

Example: blad
- publiceer, geef_uit, distribueer, hark, kauw, sterf_af
- Decide based on context of these verbs that they can be split in two clusters
  - Cluster 1: publiceer, geef_uit, distribueer
  - Cluster 2: hark, kauw, sterf_af
- => Cluster 1 becomes blad#sense1
- => Cluster 2 becomes blad#sense2

PART I: Similar words
PART II: Clustering
Intro
hierarchical vs flat
soft vs hard
clash + solutions
flat clustering (hard)
soft clustering
more about senses
Clustering Semantically Similar Words

PART I: Similar words

PART II: Clustering

Intro

hierarchical vs flat

soft vs hard

clash + solutions

flat clustering (hard)

soft clustering

more about senses

Thank you!
Hierarchical vs flat

Hierarchical
- More info
- Less efficient
- No single best algorithm, depends on task

Flat
- Less info
- More efficient
- K-means, but assumes Euclidean space, not good for nominal data (nor for probabilities?) > EM algorithm based on probabilistic models
Clustering Semantically Similar Words

Bottom-up vs top-down

- You start with every object in one cluster and start merging them.

- You start with one big cluster and look for the weakest link. The least coherent cluster is split.
Clustering Semantically Similar Words

Bottom-up vs top-down

- Splitting clusters is a clustering task in itself. Choose for bottom-up (iteration)

- However, our objects have many zero's in prob. distr. and bottom-up cannot handle that.