Topic-guided token cloud visualisations

Thomas Wielfaert, Kris Heylen, Dirk Speelman & Dirk Geeraerts

KU Leuven
Quantitative Lexicology and Variational Linguistics
Purpose of the talk

THEORETICAL

- Explore the semantic structure of a lexeme in big corpora
- Generalised: a strategy grounded in visual theory to display high-dimensional space

METHODOLOGICAL

- Interactive Scatter Plots (token clouds) in HTML5 with D3.js
- Topic models as structuring layer for big token clouds

WHAT IT IS NOT

- A WSD or WSI task evaluated against a gold standard.
Overview

1. Visualising high-dimensional space

2. Interactive token clouds

3. Topics as structuring layer
Overview

1. Visualising high-dimensional space

2. Interactive token clouds

3. Topics as structuring layer
Word similarity matrix

Black box algorithm

- Similarity matrix from a distributional semantic model: $n \times n$ dimensions
Token-level Distributional Semantic Model

One vector per token

- Bag of Words model
- PMI-weighted co-occurrence frequencies
- Hinrich Schütze’s (1998) insight that data sparsity can be overcome by moving to second-order co-occurrences
- Linguistically transparent and intuitive
## Token-level co-occurrence matrix

### Tokens of dog

<table>
<thead>
<tr>
<th></th>
<th>veterinarian</th>
<th>food</th>
<th>pet</th>
<th>loud</th>
<th>minor</th>
<th>moonlight</th>
</tr>
</thead>
<tbody>
<tr>
<td>$dog_1$</td>
<td>4.3</td>
<td>2.0</td>
<td>1.7</td>
<td>3.3</td>
<td>2.0</td>
<td>4.2</td>
</tr>
<tr>
<td>$dog_2$</td>
<td>4.2</td>
<td>1.9</td>
<td>0.9</td>
<td>2.5</td>
<td>3.2</td>
<td>2.6</td>
</tr>
<tr>
<td>$dog_3$</td>
<td>2.8</td>
<td>3.5</td>
<td>1.9</td>
<td>3.2</td>
<td>0.9</td>
<td>1.0</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$dog_n$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Similarities

Cosine or angle between vectors
Dimension reduction

Algorithms

- Standard way to deal with large similarity matrices: Kruskal’s non-metrical Multidimensional Scaling (isoMDS in R) (Cox and Cox 1999)
  Fast, but not very robust.
- More recently: T-distributed Stochastic Neighborhood Estimation (Van der Maaten 2008)
  Better for handling bigger data sets.
- Output: usually 2 or 3 coordinates per data point
Visualising coordinates

Scatter plots

Natural visualisation for data points with 2D coordinates
Amount of data you can stack in one scatter plot is limited

- Doable for fairly small (200-300 data points) samples.
- What if you want to see the bigger picture, i.e. thousands of occurrences in one plot?
Overview

1. Visualising high-dimensional space

2. Interactive token clouds

3. Topics as structuring layer
Token what?

Token clouds

- Scatter plots of similarity matrices generated by DSM
- Enriched with contextual information
- Interactive; clickable tokens
Corpora

Twente News Corpus (TwNC) (Ordelman 2002)

- 500 million words
- Newspaper material from the Netherlands.
- period: 1997–2004

Automatically lemmatized, part-of-speech tagged and syntactically parsed with the Alpino parser (van Noord 2006).
Exploring large token clouds

Case study

- Dutch polysemous word *motor*
  (n=12690 in TwNC)
  $motor_1$: engine
  $motor_2$: motorcycle
Exploring large token clouds


- overview first
- zoom and filter
- details-on-demand
Overview

1. Visualising high-dimensional space

2. Interactive token clouds

3. Topics as structuring layer
Topic models

What?

- Discovers "latent" topics using a Dirichlet prior.
- Algorithm that discovers abstract topics in a collection of documents.
- Generative model: use of joint probability distribution to label data.
Latent Dirichlet Allocation

- Most common type of topic model nowadays.
- Exists in many flavours (hLDA, HDP, sLDA, ...)
- Clusters words into topics and assigns topics to documents.
Latent Dirichlet Allocation

How?

- **words**: BoW filtered by PMI from distributional model (instead of stoplist)
- **documents**: Newspaper articles can have many topics. Pseudoparagraphs: symmetrical window of 2 sentences around the target.
- **gensim**\(^1\) package for Python

\(^1\)https://radimrehurek.com/gensim/
Topic-guided token clouds

**motor case study**

- 1) 30 topics vectors filled second-order co-occurrence frequencies from the DSM
- 2) cosine similarities
- 3) multidimensional scaling

Size depends on the number of high-loading tokens in each topic.
Topic-guided token clouds

*motor case study*

- Hover over topics to see high-loading context words, i.e. fuel, noise, car, diesel, powerful
Topic-guided token clouds

*motor* case study

- Select by clicking a topic
- Tokens that reach a threshold
Topic-guided token clouds

*motor* case study

- Zoom in on the area of interest
- See the KWIC
Future work

- Add topics or representative context words to the lowest level plots
- Develop a strategy to decide the number of topics i.e. non-parametric Hierarchical Dirichlet Process (HDP)
For more information:
http://wwwling.arts.kuleuven.be/qlvl/
thomas.wieultaert@kuleuven.be
kris.heylen@kuleuven.be
dirk.speelman@kuleuven.be
dirk.geeraerts@kuleuven.be
References 1

References 2

