Analysing Lexical Semantic Change with Contextualised Word Representations

Mario Giulianelli, Marco Del Tredici, Raquel Fernández

Institute for Logic, Language and Computation
University of Amsterdam
Number of usage types is lexeme-specific and induced from language use.

Usage vectors are characterised by contexts of occurrence – not by lists of nearest neighbouring words.

<s> ... highlighter ... </s>
**Method**

For each word of interest \( w \)

1. **extract** contextualised representations for all occurrences of \( w \) in the corpus, using a language model (e.g., BERT or ELMo)

2. **cluster** all representations of \( w \) into usage types by automatically selecting the optimal number of clusters (e.g. K-Means + silhouette score or Affinity Propagation)

3. **organise** usage clusters into diachronic usage distributions (frequency-based or probability-based)

4. **quantify** degree of change by comparing representations and usage distributions

---

**PCA visualisation of all contextualised representations for the word users as it occurs in COHA (Davies, 2012)**
Method

For each word of interest $w$

1. **extract** contextualised representations for all occurrences of $w$ in the corpus, using a language model (e.g., BERT or ELMo)

2. **cluster** all representations of $w$ into usage types by automatically selecting the optimal number of clusters (e.g. K-Means + silhouette score or Affinity Propagation)

3. **organise** usage clusters into diachronic usage distributions (frequency-based or probability-based)

4. **quantify** degree of change by comparing representations and usage distributions

Contextualised representations (left) and usage type distributions (right) for the word *users* as it occurs in COHA (Davies, 2012)
Method

For each word of interest \( w \)

1. **extract** contextualised representations for all occurrences of \( w \) in the corpus, using a language model (e.g., BERT or ELMo)

2. **cluster** all representations of \( w \) into usage types by automatically selecting the optimal number of clusters (e.g. K-Means + silhouette score or Affinity Propagation)

3. **organise** usage clusters into diachronic usage distributions (frequency-based or probability-based)

4. **quantify** degree of change by comparing representations and usage distributions

\[ 4 \]

- **Jensen-Shannon Divergence**
- **Entropy Difference**
- **Average Pairwise Distance**

between two time periods

or

average over pairs of time periods
Are the resulting usage clusters interpretable?

- literal vs metaphorical
  - ‘the ceiling of a church’
  - ‘prefer the open sky to a ceiling’
  - ‘ceiling prices’
  - ‘breaking through the ceiling’
- polysemy and homonymy
  - ‘full of questions, intensely curious’
  - ‘half fearful, half curious’
  - ‘the most curious reading’
  - ‘a curious sense of gratitude’
- entity names
  - ‘wireless device’
  - ‘wireless network’
  - ‘wirelessly’
  - ‘wirelessly’
  - ‘verizon wireless theater’
- syntactic functionality
  - ‘the refuse of the schools’
  - ‘refuse to hire’
  - ‘refuse or neglect to perform’
  - ‘refuse a draft’
  - ‘refuse, and you die’
- affixation
What types of lexical change are detected?

broadening (incl. metaphorisation): “curtain”

- I hung colored lights around my curtainless windows
- inflatable curtain-type head-protection bags
- raising the curtain on its [...] tax-reform program
- bureaucracies [...] on both sides of the curtain

narrowing: “tenure”

- employment and tenure // minority faculty in tenure
- tenure of office
- tenure-track faculty position
- reasons for short term leases and insecurity of tenure

shift: “coach”

- you can always go coach // stage coach
- cinderella - here comes your coach

new syntactic role: “download”

- to download
- a download

COHA (Davies, 2012) COCA (Davies, 2010)
Correlation with human judgements

Diachronic Usage Pair Similarity
A crowdsourced dataset of similarity judgements for more than 3K English word usage pairs (16 lemmas) from different time periods.

Data: GEMS (Gulordava & Baroni, 2011)
100 words w/ shift scores.

Shift score: average human judgement on a word’s meaning change between 1960 and 2000 (on a 4-points scale).

Metric: Spearman rank correlation between annotated change score and our three measures of change.

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>German</th>
<th>Latin</th>
<th>Swedish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency difference</td>
<td>0.068</td>
<td>0.278</td>
<td>0.276</td>
<td>0.285</td>
</tr>
<tr>
<td>Entropy difference (max)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jensen-Shannon divergence (max)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average pairwise distance (Euclidean, max)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Gulordava and Baroni (2011)
Frermann and Lapata (2016)

How similar are the two occurrences of [[federal]]?

Significant rank correlation between averaged human similarity judgements and BERT similarity scores for 10 out of 16 words.
References


