My Case, For an Adposition: Lexical Polysemy of Adpositions and Case Markers in Finnish and Latin

Daniel Chen, Mans Hulden

Creation of SNACS Corpus

- Adpositions and case markers cover a wide range of semantic fields: grounded spatial relations, abstract causal relations
- Useful to create annotation guidelines for disambiguating different senses of tokens using context of sentence

SNACS - Semantic Network for Adposition and Case Senses (Schneider et al 2018)

- Supersense hierarchy: four levels of granularity
  - Previous languages with existing SNACS corpora: English, German, Korean, Hindi, Mandarin Chinese
  - New SNACS corpus for Finnish and Latin: languages with rich case systems
  - Source texts: translations of Le Petit Prince
  - ~30 unique scene roles for each language

Construal Analysis

Finnish

<table>
<thead>
<tr>
<th>Sentence (from Pikku Pinssi, Chapter IV)</th>
<th>Scene Role</th>
<th>Function Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Ystäväni lähti pois lampuis-e-en friend-1P.POSS went away sheep-LOCATIVE-3P.POSS</td>
<td>ANCILLARY</td>
<td>ANCILLARY</td>
</tr>
<tr>
<td>b. En ehdyn myös vähän pituude-ssa, I err also a little length-NESSIVE</td>
<td>TOPIC</td>
<td>LOCUS</td>
</tr>
</tbody>
</table>

Latin

<table>
<thead>
<tr>
<th>Sentence (from Regulus, Chapter IV)</th>
<th>Scene Role</th>
<th>Function Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Quo sit omnium curn exi With them.ACT</td>
<td>RECIPIENT</td>
<td>ANCILLARY</td>
</tr>
<tr>
<td>b. meh ellei admiration-em maga-un move non poter me.DATIVE-admiration-3P ACC great-3P ACC he move not was_able-3P PST ... he was not able to move great admiration for me.</td>
<td>EXPERIENCER</td>
<td>GOAL</td>
</tr>
</tbody>
</table>

K-Means Clustering Analysis

- All adpositions and case-marked words are represented by Multilingual BERT (BERT base, uncased) embeddings generated from automatically tokenized WordPieces.
  
**In tensor([0.0718, -4.9129, 0.1489, 2.4959, -3.2524])**
  
- Concatenate the last 4 hidden layers => each WordPiece token has 3,072 dimensions.
- Achieves higher cosine similarity between tokens
- Use principal component analysis (PCA) to reduce dimensionality to 100 dimensions, for more efficient processing.

- **K-Means Clustering Algorithm:** identifies k (user-specified) cluster centers (prototype / mean of data points), assigns each data point to a cluster.

Results + Analysis

1. Adpositions cluster well together, since they are separate words that typically make up the entirety of their own WordPiece token.

2. Adpositions are found in separate clusters from their case-marked noun objects => mBERT identifies no semantic similarity between them.

3. Multword adpositions are not found in the same cluster.

4. Clusters do not group morphological alternations of the same case marker together.

- Finnish: elative case markers -sta and –stanne are grouped in different clusters.
- Latin: first declension ablative singular -ia is with other ablative singular endings, but second declension ablative singular –ulo is with noun stems.

=> mBERT cluster partitions are strictly orthographical. Can’t quite begin to account for semantic variation when morphological variation is not accounted for.