K-means Clustering for POS Tagger Improvement

Gabi Rolih

Department of Linguistics and Philology, Uppsala University
Engelska parken, Thunbergsv. 3H, 751 26 Uppsala
gabirolih@gmail.com
Ljubešić, Erjavec and Fišer (2017):
*Adapting a State-of-the-Art Tagger for South Slavic Languages to Non-Standard Text*

- Efficiently using Brown clustering information to improve ReLDI tagger

**Project:** Using K-means clustering to improve the ReLDI tagger and compare with Brown clustering
Previous Work

- Turian et al, 2010
  - Compare Brown clustering, Collobert and Weston embeddings, HLBL embeddings for NER tasks
  - Brown clusters show highest accuracy
- Owoputi et al, 2013
  - Use Brown clusters to improve PoS tagging in informal conversational texts
- Lin and Wu, 2009
  - Use K-means clustering on phrases for NER and query classification with great results
• Clustering: SIWaC v2.0 web corpus of Slovene (1.2 billion tokens)
• Tagger: Janes-Tag v1.2 annotated dataset
  – Slovene CMC texts: forum posts, tweets, comments
  – Training: 60,367 tokens
  – Testing: 7,484
K-means Clustering

- $K =$ number of clusters = number of centroids
- Random initialization of centroids
- In each iteration:
  1. Assign clusters
  2. Move centroids
- Repeat until convergence

Source: DATASCIENCE.COM: Introduction to K-means Clustering
Word2Vec

- Converts words to vectors based on their context
- Single layer of a feed-forward neural network
- Probability of word co-occurring with other words
- Output: a feature matrix of words
Clustering settings

- **Word2Vec**: Gensim library
  - Only words with frequency $> 50$
  - Window size is 2
- **K-Means**: Scikit-learn package
  - $K = 2000$
<table>
<thead>
<tr>
<th>Method</th>
<th>ReLDI trained on CMC data</th>
<th>Brown</th>
<th>K-means</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSD</td>
<td>84.15</td>
<td>85.17</td>
<td>88.32</td>
</tr>
<tr>
<td>PoS</td>
<td>89.85</td>
<td>91.12</td>
<td>92.88</td>
</tr>
</tbody>
</table>
Conclusions

• Clustering information improves tagger accuracy
• K-means combined with Word2Vec outperforms Brown
• Future work:
  – Finding a more efficient way of including K-means data into tagger
  – Testing of other parameter settings
  – Exploration of other clustering techniques