Typological Feature Prediction with Matrix Completion

Annebeth Buis
anne.buis@colorado.edu
Mans Hulden
mans.hulden@colorado.edu

Motivation

- Typological features are informative to many cross-lingual tasks in NLP. Unfortunately typological data is often sparse or incomplete and generating it is costly. For example, in the World Atlas of Language Structures (WALS) the value of 80% of language-feature combinations is undefined.

WALS and Preprocessing

- WALS contains features in 11 domains: phonology, sign languages, morphology, nominal categories, nominal syntax, verbal categories, word order, simple clauses, complex sentences, lexicon and other. Meta-features (e.g., isocodes, language family, genus, etc.) were not included in the experiments.
- The original WALS matrix contains categorical feature values, which were binarized before running matrix completion. We excluded 214 languages for which only 1 feature value has been recorded in WALS.
- No additional preprocessing or excluded features.

Matrix Completion

- Matrix completion methods have been used extensively with sparse matrices and are able to learn more holistic patterns in the data than individual local predictors (such as our logistic regression baseline).
- IterativeSVD (based on Troyanskaya et al., 2001) learns a low-rank approximation of the original matrix by using Singular Value Decomposition (SVD).

Experiments

- Leave-one-out-cross-validation to predict each language × feature-combination that is currently in WALS.
- Results are compared against a majority class baseline and a logistic regression classifier.
- To test for robustness, 2 additional experiments:
  1. Leaving out features in the same domain (e.g., Phonology, Morphology, etc.)
  2. Leaving out languages with the same language family (e.g., Indo-European, Afro-Asiatic, etc.)

Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Georgi et al. (2010)</td>
<td>65.5%</td>
<td>Language clustering</td>
</tr>
<tr>
<td>Takamura et al. (2016)</td>
<td>75.5%</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>Murawaki (2017)</td>
<td>73.0%</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>Baseline 1</td>
<td>53.1%</td>
<td>Majority class</td>
</tr>
<tr>
<td>Baseline 2</td>
<td>65.7%</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>Matrix completion</td>
<td>74.3%</td>
<td>IterativeSVD</td>
</tr>
<tr>
<td>without domain</td>
<td>61.6%</td>
<td>IterativeSVD</td>
</tr>
<tr>
<td>without language family</td>
<td>71.2%</td>
<td>IterativeSVD</td>
</tr>
</tbody>
</table>

- Prediction accuracy per language does not improve with more examples.
- Feature accuracy improves when more examples are available.

Conclusion

- Matrix completion outperforms the baselines on the WALS data and performs on par with previous work.
- Matrix completion requires minimal preprocessing and can easily be used with any typological database.
- Our work has shown that treating WALS as a matrix is an effective approach. This should be further explored in future work.

Data and code are available at: github.com/annebeth/wals-matrix-completion