Ensemble Self-Training for Low-Resource Languages: Grapheme-to-Phoneme Conversion and Morphological Inflection

Xiang Yu  
Ngoc Thang Vu  
Jonas Kuhn  
IMS, University of Stuttgart, Germany  
{xiangyu,thangvu,jonsakuhn}@ims.uni-stuttgart.de

Summary

- Iterative ensemble optimization and data augmentation
- Based on large amount of diverse simple models
- Effective for low-resource scenarios
- 1st in the grapheme-to-phoneme conversion task
- 4th in the morphological inflection task

Ensemble Self-Training

General Workflow

1. function EnsembleSelfTraining(L, U, T)  
   Require: labeled data L, unlabeled data U, model types T
2. Initial data L_0 = L
3. Model pool M = ∅
4. for n = 0, 1, ..., N do
5.   for t = 1, 2, ..., T do
6.     m_t = TRAIN(L_t, L_0)  
7.     M = M ∪ {m_t}
8.   end for
9.   E = SEARCHENSEMBLE(M)  
10. Sample ū ∼ U
11. l = SELECTDATA(E, ū)  
12. L_n+1 = AGGREGATEDATA(L_n, l)
13. U = U ∪ {l}
14. end for
15. return E, L_N
16. end function

Ensemble Search

- Search for the optimal ensemble with genetic algorithms
- A binary code to represent an ensemble, e.g. 0101001110
- Fitness of an ensemble is the accuracy on the dev set
- Use selection, crossover, and mutation to evolve the ensemble

Data Selection

- Use the ensemble to predict a batch of unlabeled data
- Select the data with high agreement with the ensemble
- Add as new training data for the next iteration

Task & Data

- Task 1 of SIGMORPHON Shared Task
- Map a sequence of graphemes to a sequence of phonemes
  e.g. excusær → eskýze
- Unlabeled data: word lists mostly extracted from OpenSubtitles

Models

- 4 types of models:
  1. baseline pair n-gram model
  2. seq2seq model with attention
  3. seq2seq model with hard monotonic attention
  4. hybrid seq2seq/tagging model: predict a short sequence for each input character
- Paired with l2r and r2l generation directions and 2 random seeds

Results

<table>
<thead>
<tr>
<th>Model</th>
<th>WER</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMS</td>
<td>13.81</td>
<td>2.76</td>
</tr>
<tr>
<td>CLUZH</td>
<td>14.13</td>
<td>2.82</td>
</tr>
<tr>
<td>DeepSPIN-3</td>
<td>14.15</td>
<td>2.92</td>
</tr>
<tr>
<td>Pair n-gram</td>
<td>22.00</td>
<td>4.97</td>
</tr>
<tr>
<td>LSTM</td>
<td>16.84</td>
<td>3.99</td>
</tr>
<tr>
<td>Transformer</td>
<td>17.51</td>
<td>4.30</td>
</tr>
</tbody>
</table>

Table: Average word error rates (WER) and phone error rates (PER) on test set.

Analysis

- Analyze the contribution of each factor:
  - Ensemble much better than single models
  - Lower model diversity (only hybrid model) leads to lower ensemble performance despite higher average model performance
  - Worse performance with data augmentation
- Simulate low-resource scenario:
  - 200 training instances for each language
  - Better performance with data augmentation

Morphological Inflection

- Task 0 of SIGMORPHON Shared Task
- Generate inflected word form from lemma and morphological features
  e.g. jagen → V.SBJV.PL;3;PST → jagten
- Unlabeled data: recombine the lemma and morphological features

Models

- 2 Types of models:
  1. seq2seq model with soft attention
  2. seq2seq model with hard monotonic attention
- Paired with l2r and r2l generation directions and 2 random seeds

Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>CU-Ling-01-0</td>
<td>0.912</td>
</tr>
<tr>
<td>DeepSPIN-02-1</td>
<td>0.909</td>
</tr>
<tr>
<td>UIUC-01-0</td>
<td>0.905</td>
</tr>
<tr>
<td>IMS-00-0</td>
<td>0.892</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.858</td>
</tr>
<tr>
<td>LSTM-Aug</td>
<td>0.888</td>
</tr>
<tr>
<td>Transformer</td>
<td>0.901</td>
</tr>
<tr>
<td>Transformer-Aug</td>
<td>0.903</td>
</tr>
</tbody>
</table>

Table: Average accuracy on test set.

Analysis

Figure: Performance difference between our system and Transformer-Aug wrt. training data size.

- Our system performs relatively better in low-resource scenarios
- No clear relation between performance and language family

Acknowledgments

This work was in part supported by funding from the Ministry of Science, Research and the Arts of the State of Baden-Württemberg (MWK), within the CLARIN-D research project.

References


