Improving a simple bigram HMM part-of-speech tagger by latent annotation and self-training

Zhongqiang Huang¹, Vladimir Eidelman¹, Mary Harper¹,²
¹University of Maryland, College Park
²Johns Hopkins Human Language Technology Center of Excellence

Introduction
- Part-of-speech (POS) tags may be too course to adequately model tagging.
- Trigram tagger, with weaker independence assumptions than bigram tagger, can be viewed as explicit annotation of each tag with its preceding tag.
- We automatically split observable POS tags into sub-tags.
- Tag specialization may provide useful distinctions
- Better fit to data

Goals:
- Develop and evaluate a Latent Annotation (LA) tagger for Chinese that automatically specializes tags
- Use Self-training → more accurate models

What's different with latent annotations?

<table>
<thead>
<tr>
<th></th>
<th>Standard bigram tagger</th>
<th>Bigram tagger with LA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training likelihood</td>
<td>( p(w_i^c, c^*) = \prod_i p(c_i</td>
<td>x_i) p(w_i</td>
</tr>
<tr>
<td>Parameters</td>
<td>( p(w_i</td>
<td>a_i, b_i) )</td>
</tr>
<tr>
<td>Markov order</td>
<td>1st order on POS tags</td>
<td>1st order on latent POS tags not on POS tags themselves</td>
</tr>
<tr>
<td>Parameter estimation</td>
<td>Count, smooth &amp; normalize</td>
<td>Iterative EM</td>
</tr>
<tr>
<td>Tagging</td>
<td>Viterbi decoding</td>
<td>Approximated decoding</td>
</tr>
</tbody>
</table>

- **Goal:** learns subcategories of POS tags to model different types of words
- **Bad:** distracts model from learning common phenomena when strongly discriminating among rare words
- **Solution:** force equality for rare words

Experimental setup
- CTB6-unlabeled Chinese newswire data
- Signals: Train Dev Test Unlabeled
  - Sentences: 24k 2k 2k 210k
  - Words: 679k 51k 53k 6,255k

- Evaluated three taggers
  - Bigram
  - Trigram
  - Bigram-LA

- Trained on 10%, 20%, 40%, 60%, 80%, and 100% of the labeled training data
- All data is pre-processed by UW Decatur
- All three taggers use a character-based method for handling unseen Chinese words

Performance of the taggers on the test set

<table>
<thead>
<tr>
<th>Tagger</th>
<th>Token accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bigram</td>
<td>92.25</td>
</tr>
<tr>
<td>Trigram</td>
<td>93.99</td>
</tr>
<tr>
<td>Bigram+LA</td>
<td>94.53</td>
</tr>
<tr>
<td>Bigram+LA+ST</td>
<td>94.78</td>
</tr>
</tbody>
</table>

Results

Learning curves of LA tagger with & without rare word handling

Conclusions
- LA tagger: Improves POS tagging accuracy
- More benefit from self-training than standard n-gram taggers
- Using less labeled data + self-training
- Rare word handling important

Next steps:
- Cross-lingual study
- Evaluate benefit of more data and data selection methods
- Investigate other ways to use unlabeled data