Combining Statistical Models for POS Tagging using Finite-State Calculus

Miikka Silfverberg and Krister Lindén

Department of Modern Languages
University of Helsinki

The 18th Nordic Conference of Computational Linguistics
POS Tagging

- POS tagging is a core task in language technology.
- POS taggers are traditionally constructed using statistical methods or rule based methods.
- We are working to create a framework which can use different kinds of statistical models together with rules.
- We demonstrate our framework by constructing an HMM augmented with contextual lexical models.
Problems with Statistical POS Taggers

- Sufficiently large tagged corpora exist for few languages.
- Statistical models are domain specific.
- The syntax and morphology of a lot of languages differ considerably from those of English.
Incorporating a Variety of Knowledge Sources using WFSTs

WFSTs can describe a large class of weighted sequential models including statistical models such as HMMs and linguistic rules like Constraint Grammar rules.

POS tagging can be accomplished using WFST algorithms.
Like a typical HMM, our taggers consist of a lexical model and contextual models.

For each word $w$ and tag $t$, the lexical model gives a probability estimate $P(w|t)$ based on unigram probabilities.

The lexical model includes a lexicon for known words and a number of suffix-based guessers for unknown words.
The tag profiles of the words in the sentence are compiled into a sentence transducer and re-scored using contextual models.

There is no limitation of the kind of contextual models that can be used. Both statistical and rule based models can be combined.
The task of the lexical model is to give an estimate for the unigram probability $P(w|t)$.

For known words, we simply use unigram probabilities $N(w, t)/N(t)$ computed from a training corpus.

For unknown words, we need to estimate the probabilities using suffixes of the word. We use $N(s_i, t)/N(t)$ for different lengths of suffixes $i$ and compute the probability $P(w|t)$ as a weighted sum of these.

We use different guessers for upper case and lower case words as well as sentence initial words.
Tag N-Gram Models

- One part of the current model emulates a traditional second order HMM.
- We implement each of the n-gram models as a set of \( n \) transducers.
- E.g. the trigram model consist of 3 transducers \( M_0, M_1 \) and \( M_2 \), which assign probability \( P(t_i|t_{i-1}, t_{i-2}) \) to trigrams starting at positions \( 3n, 3n + 1 \) and \( 3n + 2 \) in the sentence.

<table>
<thead>
<tr>
<th></th>
<th>Mr.</th>
<th>Gelbart</th>
<th>also</th>
<th>has</th>
<th>fun</th>
<th>with</th>
<th>language</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M_0 )</td>
<td>NNP</td>
<td>NNP</td>
<td>RB</td>
<td>VBZ</td>
<td>NN</td>
<td>IN</td>
<td>NN</td>
</tr>
<tr>
<td>( M_1 )</td>
<td>NNP</td>
<td>NNP</td>
<td>RB</td>
<td>VBZ</td>
<td>NN</td>
<td>IN</td>
<td>NN</td>
</tr>
<tr>
<td>( M_2 )</td>
<td>NNP</td>
<td>NNP</td>
<td>RB</td>
<td>VBZ</td>
<td>NN</td>
<td>IN</td>
<td>NN</td>
</tr>
</tbody>
</table>

Mr. Gelbart also has fun with language .
One of the weaknesses of HMMs is that lexical data is isolated from contextual data.

Improvement in accuracy can be gained by conditioning the possible tags of a given word according to the tags surrounding it.

We use a model which estimates probabilities $P(w_i|t_{i-1}, t_i, t_{i+1})$. 
Context Dependent Lexical Models

To cope with previously unseen combinations of words and tag trigrams, we also estimate probabilities $P(w_i|t_i, t_{i+1})$ and $P(w_i|t_{i-1}, t_i)$.

The context dependent lexical model is only used for known words.
Data

- We used the WSJ corpus with Collins splits for the English tagger.
- For Finnish we used morphologically analyzed and disambiguated newspaper text.

### Token counts

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>Finnish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>969905</td>
<td>1027514</td>
</tr>
<tr>
<td>Tuning</td>
<td>148158</td>
<td>181437</td>
</tr>
<tr>
<td>Testing</td>
<td>171138 (2.43%)</td>
<td>156572 (10.41%)</td>
</tr>
</tbody>
</table>
The tag count for English is higher than the number of simple POS tags in the Penn Treebank POS tag set, because of compound tags.

<table>
<thead>
<tr>
<th>POS Tags</th>
<th>English</th>
<th>Finnish</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>81</td>
<td>776</td>
</tr>
</tbody>
</table>
The lexical model for Finnish

There were a lot of sentence initial unknown words in the Finnish material.

Both using the lower case and upper case guesser for these words yielded poor results, so we decided to train a separate guesser for sentence initial words.

For English guessers give an estimate $P(w|t)$ for every tag given the unknown word $w$ and all of the guesses are included in the sentence transducer. For Finnish we only use the 10 best guesses, which improves accuracy.
## Results

<table>
<thead>
<tr>
<th>Language</th>
<th>Seen</th>
<th>Unseen</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>English</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TNT</td>
<td>96.77%</td>
<td>85.19%</td>
<td>96.46%</td>
</tr>
<tr>
<td>Hunpos</td>
<td>96.88%</td>
<td>86.13%</td>
<td>96.58%</td>
</tr>
<tr>
<td>Hfst</td>
<td>97.13%</td>
<td>83.72%</td>
<td>96.77%</td>
</tr>
<tr>
<td><strong>Finnish</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hunpos</td>
<td>98.06%</td>
<td>76.83%</td>
<td>95.62%</td>
</tr>
<tr>
<td>Hfst</td>
<td>97.98%</td>
<td>81.04%</td>
<td>96.02%</td>
</tr>
</tbody>
</table>
Conclusions

- We have demonstrated a framework for constructing taggers, which makes it possible to use a variety of statistical models in parallel.
- We achieved improved accuracy compared to two other HMM-based taggers.
- Tagging Finnish shows that a good framework for tagging should be flexible so that new models, such as the sentence initial guesser can be added.
Future Work

- We would like to add rules, which requires a rule compiler. This would reduce systematic errors in the tagger.
- Applying the framework to other sequential labeling tasks such as tokenization, chunking and named entity recognition.
- Improving the guessers for unknown words.
Combining Statistical Models for POS Tagging using Finite-State Calculus

Silfverberg & Lindén

Introduction

POS Taggers using WFSTs

Constructing Taggers for Finnish and English

Evaluation

Conclusions and Future Work

Thank you!