Part-of-Speech Tagging with Finite-State Morphology

Bryan Jurish
jurish@ling.uni-potsdam.de

Abstract

Part-of-Speech (PoS) Tagging is the identification for each input token of its lexical category. Traditional tagging techniques such as Hidden Markov Models (HMMs) make use of both lexical and bigram probabilities derived from a tagged training corpus in order to compute the most likely PoS tag sequence for each input sentence. By allowing use of a finite-state morphology component, the dwdst PoS tagging library extends traditional HMM techniques by the inclusion of lexical class probabilities and theoretically motivated search space reduction.

The morphological component is implemented as a finite-state transducer $T_{\text{morph}} : \Sigma^* \rightarrow 2^{(T \cup F)}$, which maps tokens $w_i$ to analysis sets $A_i \subseteq (T \cup F)$, themselves encoded as transducers: $A_i \equiv \pi_2(T_{\text{morph}}(w_i))$.

Viterbi Algorithm adapted to compute the most probable sequence of PoS tags $t_{1..n}$ for an input sequence $w_{1..n}$ with PoS classes $C_{1..n}$.

Given the most likely sequence of PoS tags $t_{1..n}$, the output morphological analysis sets $A'_i$ are restricted by inverting the extracted classes $C_i$: $A'_i = C_i^{-1}(\{t_i\})$.

Results

<table>
<thead>
<tr>
<th>Tag Set</th>
<th>Feature Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Superset $T$ of the 56-tag Stuttgart-Tübingen Tag Set (STTS)</td>
<td>39 feature types and 566 feature values $F$ currently implemented</td>
</tr>
</tbody>
</table>

The morphological component is implemented as a finite-state transducer $T_{\text{morph}} : \Sigma^* \rightarrow 2^{(T \cup F)}$, which maps tokens $w_i$ to analysis sets $A_i \subseteq (T \cup F)$, themselves encoded as transducers: $A_i \equiv \pi_2(T_{\text{morph}}(w_i))$.

Viterbi Algorithm adapted to compute the most probable sequence of PoS tags $t_{1..n}$ for an input sequence $w_{1..n}$ with PoS classes $C_{1..n}$.

Given the most likely sequence of PoS tags $t_{1..n}$, the output morphological analysis sets $A'_i$ are restricted by inverting the extracted classes $C_i$: $A'_i = C_i^{-1}(\{t_i\})$.

Conclusion

The use of lexical class probabilities in addition to traditional raw lexical probabilities resulted in a 17.6% reduction in errors for the corpus configuration given above. The linguistically motivated search space reduction provides a 94.4% improvement in speed for the German morphological component.