Kollokationen im Wörterbuch

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*.

Input:	Linda	Nakew	wird	die	Mannschaft	verstärken	.
	Linda	*Nakew*	*will*	*the*	*team*	*join*	.
Morphological Analysis:	$\{\texttt{NE}.first, \texttt{NE}.last\}$	$\emptyset$	$\{\texttt{VAFIN}.3rd.sg.pres, \texttt{VVFIN}.3rd.sg.pres, \texttt{VVIMP}.sg\}$	$\{\texttt{ART}.sg.nom.fem, \vdots, \texttt{PDS}.nom.sg.fem, \vdots, \texttt{PRELS}.acc.pl\}$	$\{\texttt{NN}.masc.sg.nom, \vdots, \texttt{NN}.fem.sg.*\}$	$\{\texttt{VVFIN}.1st.pl.pres, \vdots, \texttt{VVINF}\}$	$\{\$.\}$
PoS Extraction:	$\{\texttt{NE}\}$	$\{\texttt{ART}, \vdots, \texttt{XY}\}$	$\{\texttt{VVFIN}, \texttt{VVIMP}, \texttt{VAFIN}\}$	$\{\texttt{ART}, \texttt{PDS}, \texttt{PRELS}\}$	$\{\texttt{NN}\}$	$\{\texttt{VVFIN}, \texttt{VVINF}\}$	$\{\$.\}$
Disambiguation:	NE	NE	VAFIN	ART	NN	VVINF	$.
PoS Restriction:	$\{\texttt{NE}.first, \texttt{NE}.last\}$	$\{\texttt{NE}\}$	$\{\texttt{VAFIN}.3rd.sg.pres\}$	$\{\texttt{ART}.sg.nom.fem, \texttt{ART}.sg.acc.fem, \texttt{ART}.pl.nom.*, \texttt{ART}.pl.acc.*\}$	$\{\texttt{NN}.masc.sg.nom, \vdots, \texttt{NN}.fem.sg.*\}$	$\{\texttt{VVINF}\}$	$\{\$.\}$

Tag Set

Superset T of the 56-tag *Stuttgart-Tübingen Tag Set* (STTS)

Feature Set

39 feature types and 566 feature values F currently implemented

Morphological Analysis

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

$$A_i = \pi_2(T_{morph}(w_i))$$

PoS Extraction

To accomodate varying morphological conventions, an additional transducer $T_{tagx} : 2^{(T \cup F)^*} \rightarrow 2^T$ may be employed to map analysis sets A_i to lexical classes $C_i : \pi_1(A_i) \rightarrow 2^T$

$$C_i = A_i \circ T_{tagx}$$

HMM Disambiguation

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 $C_{1..n}$

$$Dis(w_{1..n}) = \arg\max_{t_{1..n}} \prod_{i=1}^{n} P(t_i|t_{i-1}) P(w_i|t_i) P(\pi_2(C_i)|t_i)$$

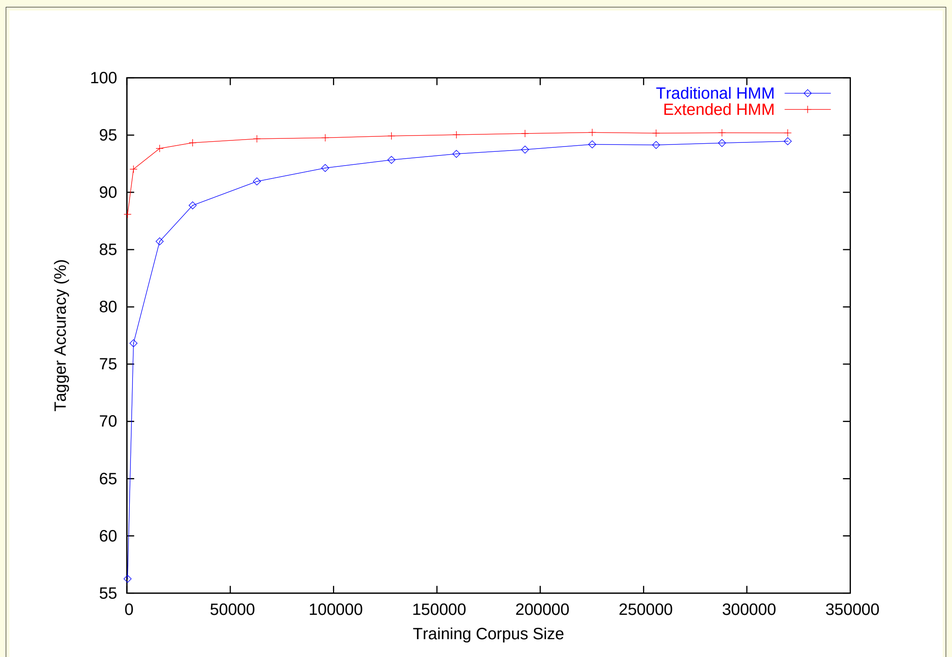
PoS Restriction

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

Corpus (NEGRA)	Base Size:	355096 tok
	Training Size:	319764 tok
Morphology	Recognition:	97.21%
	Coverage:	95.13%
	Class Size:	3.12 tags
Disambiguation	Saves:	61.13%
	Success:	97.67%
Global	Accuracy:	95.19%
	Throughput:	25047 tok/s



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.

Alexander von Humboldt Stiftung/Foundation
Zukunfts-investitions-programm zip
Berlin-Brandenburgische Akademie der Wissenschaften