Part-of-Speech Tagging (with Markov Models)

Simon Clematide
Institute of Computational Linguistics
University of Zurich

http://www.cl.unizh.ch/siclemat/talks/markov/

Synopsis

Program

- Part-of-Speech tagging as a classification problem
  - How hard is it?
- Applying HMM to PoS
  - Interpretation of HMMs for PoS-Tagging
  - Finding the best tag sequence (aka. Viterbi algorithm)
- Empirical problems
  - Unknown words & zero frequency bigrams
  - Guessing & Smoothing
- State-of-the-Art HMM trigram tagger
  - Comparison of HMM techniques
  - tagging quality with respect to training size

The Problem

Given a sequence of words classify them according to a tag set.

NNP	VBZ	NN	NNS	CD	NN		
Fed	*raises*	*interest*	*rates*	*0.5*	*%*	*in*	*effort*
						to	*control*
							inflation

- Different tag set sizes
  - From dozens (Penn: 45; Brown: 87) to few hundred tags (London-Lund 197)
- Different information included
  - Traditional word categories (but including interpunctuation elements)
  - (Distributional) syntactic information
  - Morphological (*singular vs. plural, case*)
  - Semantic (*proper names, pronominal vs. attributive readings*)
  - Tags for foreign language inclusions
- Often hierarchically organized

Hard Problem or Not?

Lexical ambiguity is crucial

- How many tags are known or possible for a word?

		VB			
	VBZ	VBP	VBZ		
NNP	NNS	NN	NNS	CD	NN
Fed	*raises*	*interest*	*rates*	*0.5*	*%*

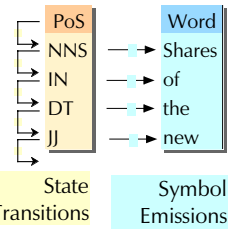
Different information sources

- **Sequence of words** is not very helpful (surprisingly)
- **Sequence of tags** is better
- **Empirical frequency of word/tag coocurrence** is effective
  - Tagging a word with its most frequent PoS gets 90+% accuracy.
  - Baseline tagger for English

Applying Markov Models to Tagging

The tagging task interpretation for HMMs

- (Hidden) States = Part-of-Speech Tags
  (except special start and end state)
- Emission Symbols = Tokens
- Transition Probabilities = Contextual Probabilities
- Emission Probabilities = Lexical Probabilites
  - Important: What is the probability of tag *t* generating word *w* ?
  - And **not**: What is the probability of word *w* of having tag *t*?
- Our Task III of HMMs = Find the most probable tag sequence for a given sequence of tokens!
  - Tagging is a case for Viterbi!



$$\arg\max_{S \in \mathcal{S}} P(W,S)$$

Empirical Problems

HMMs built by pure MLE parameter estimation from a training corpus have problem with new material.

- **Unknown words** have zero lexical probabilities. **Zero propagation** of multiplication gives probability 0 for every sequence.

$$P(w_u \mid s) = \frac{freq(w_u, s)}{freq(s)} = \frac{0}{freq(s)} = 0$$

$$P(W,S) = \left(\prod_{i=1}^{i=t} P(s_i \mid s_{i-1}) \underbrace{P(w_i \mid s_i)}_{0 \text{ for } w_u} \right) \cdot P(s_e \mid s_t)$$

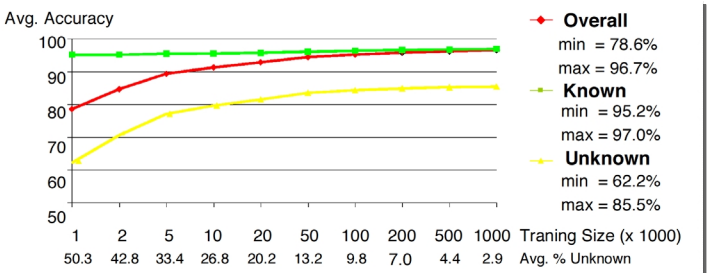
- Same Problem with **unseen bigrams** (tag sequences)
  - Regardless of the quality of the surrounding tags or lexical probabilities!
- Solution (see M&S for various techniques)
  - **Guess** the tag of unknown words!

Naive vs. State-of-the-Art Tagging

- Naive approach:
  - Bigram model (first order HMM)
  - Smoothing: addition of $c = 0.5$ to zero frequencies
  - Unknown words: tag distribution estimated from all words
  - No sentence boundary and capitalization info
- State–of–the–Art:
  - Trigram model (second order HMM)
  - Smoothing: Context–independent linear interpolation
  - Unknown words: suffix analysis and successive abstraction
  - Automatic modeling of sentence boundaries and capitalization

	naive	state–of–the–art	Δ
WSJ (English)	95.0%	96.7%	+1.7%
NEGRA (German)	92.4%	96.7%	+4.3%

Tagging Accuracy and Training Size



Penn Treebank: 1,2 million tokens newspaper text (Wall Street Journal)

randomly selected training (variable size) and test parts (100,000 tokens)

10 iterations for each training size; training and test parts are disjoint.

No other sources were used for training.