Speech Recognition Project Presentation: POS Tagging with WFSTs

Jose Cambronero

05/13/2015

Agenda

- Brief intro to POS tagging and approaches
- FSTs in POS tagging in literature: some interesting examples
- Our setup and approach
- Results
- Possible improvements

Brief Intro to POS tagging

- DET, N, V, ADJ ...
- We want a tagger F, s.t. F(this is an example) $\rightarrow$ DT BEZ AT NN
- Succinct definition provided by Jurafsky: $argmax_{t_0^n} P(t_0^n | w_0^n)$, where t_i corresponds to tag at position i, w_i corresponds to word at position i in sentence, and n is the last index in our sentence.
- Variety of approaches: generative vs. discriminative

Tzoukermann and Radev: *Use of Weighted Finite State Transducers in Part of Speech Tagging* (1997)

- One of earliest papers on use of FSTs for POS tagging
- Cascade of FSTs: tokenization, morphological analysis, linguistic disambiguation, statistical disambiguation
- Use of unknown tag at every stage to avoid failures in composition, simply associates highest penalty with it, so only used if nothing else available
- Concept of genotypes: possible POS tags, words mapped into genotypes and distributions calculated over the latter
- Incorporates linguistic information directly through negative constraint transducers: works by increasing the cost of a given path
- trained on 76162 tokens and tested on a separate 2200 tokens
- Results: 92.1% 1-grams, 93.4% 2-gram extension and 96% when extending with negative constraints and 3-grams

Kempe: *Part-of-Speech Tagging with Two Sequential Transducers* (2001)

- ambiguity classes (similar to Tzoukermann/Radev's genotypes): c_i, r_i, and tags t_i
- ambiguity class mapper/guesser + 2 sequential transducers T_1, T_2
  - $T_1 : c_i \rightarrow r_i$ (left to right)
  - $T_2 : r_i \rightarrow t_i$ (right to left)
- Construction of FSTs reflects probabilities, but FSTs themselves have no weights
- Accuracy: 96.67% (tags 45600 words per second)

Silfverberg and Lindén: *Part-of-Speech Tagging using Parallel Weighted Finite-State Transducers* (2010)

- 4 WFSTs: weighted lexicon (word to lemma and POS) and guesser in first stage (provide 5 best possible POS tags for each word), Q_o and Q_e in second (rescoring) stage
  - POS guesser for unknown words:assigns POS distribution P from the set of known words W, where each $w \in W$ is a word that has the longest common suffix with the unknown word.
  - Q_o: bigram model applied at odd positions
  - Q_e: bigram model applied at even positions
- Weighted intersecting composition: simulates intersecting with larger Q but more performant
- includes word lemmas in bigram models
- Europarl corpus: 98.29% accuracy for English using bigrams
  - Europarl is more homogeneous than other corpora (as per authors)

Our approach and setup

- An HMM approach: $P(w_i \mid t_i) * P(t_i \mid h)$ with a variety of differing history orders
- OOV strategy
  - Frequency threshold based
  - simple vs clustering
- Unseen tag sequences
  - interpolated smoothing approaches

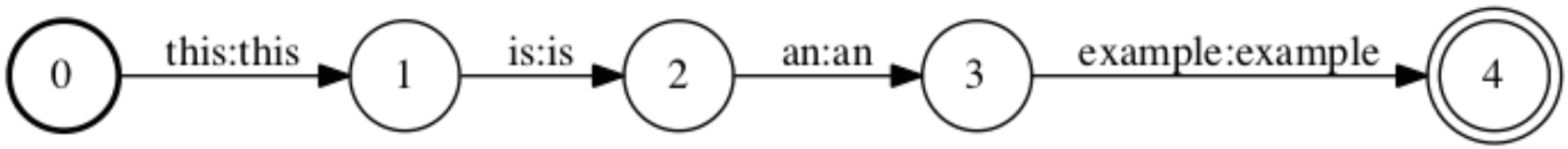
Our approach and setup (continued)

- T: a WFST from token to possible POS

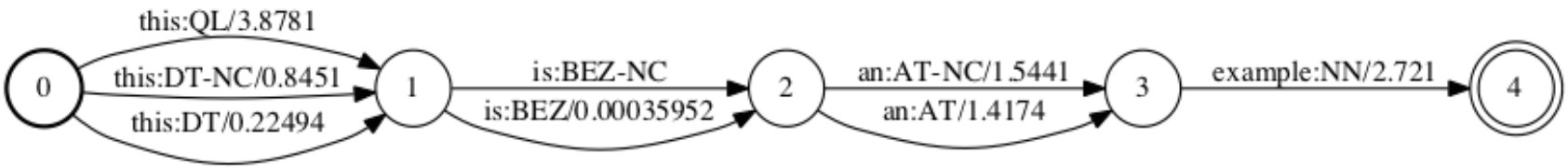
- H: a WFST n-gram rescoring model
- All probabilities mapped to $-log$ space to convert to penalties
- Predicted POS tag sequence corresponds to: shortestPath$((S \circ T) \circ H))$

An Example

- FST of our sentence *this is an example*

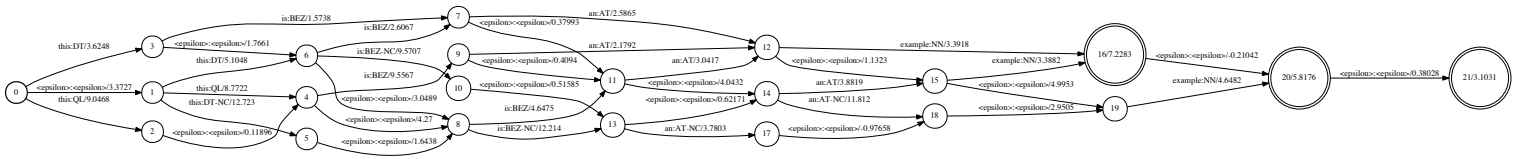


- $S \circ T$

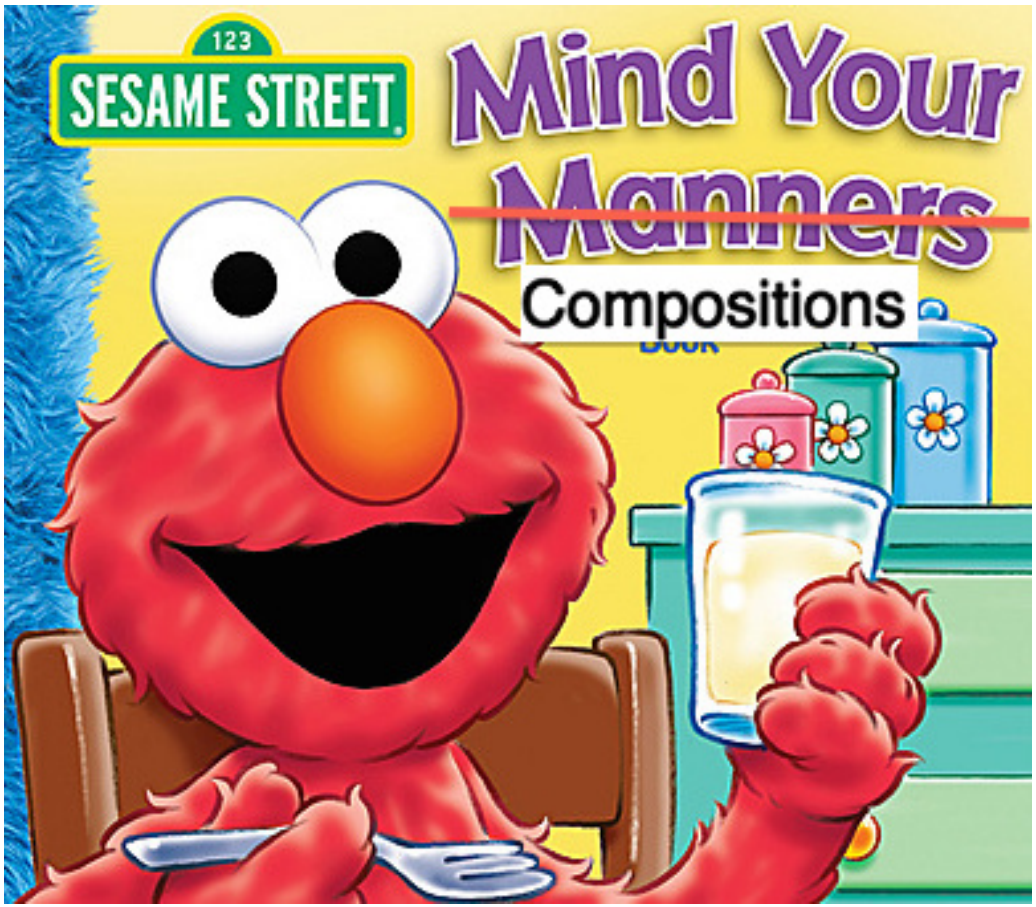


An Example (continued)

$(S \circ T) \circ H$ over which we can calculate the shortest path to obtain the optimal tag sequence



Mind your compositions!



Mind your compositions!

- composition is associative so $(S \circ T) \circ H \equiv S \circ (T \circ H)$
- The latter will blow up and your tagging will be non-performant ... if it finishes!

N-gram order	$(S \circ T) \circ H$	$S \circ (T \circ H)$
2	0.02	0.35
3	0.03	6.36
4	0.07	$T \circ H$ didn't terminate in reasonable time

Figure 1: Seconds to tag *this is an example*

Data

- Brown corpus
- Train on 80%, test on remaining

Results: Simple OOV strategy

Model	TH 6 Accuracy	TH 3 Accuracy
2-gram Katz	90.85%	92.06%
2-gram Kneser Ney	91.03%	92.23%
2-gram Witten Bell	91.02%	92.23%
3-gram Katz	91.59%	92.71%
3-gram Kneser Ney	91.90%	92.97%
3-gram Witten Bell	91.86%	92.91%
4-gram Katz	91.00%	92.26%
4-gram Kneser Ney	91.62%	92.68%
4-gram Witten Bell	91.50%	92.58%

Figure 2: Test accuracy using simple OOV strategy

Clusters in OOV clustering strategy

Unknown cluster	criteria
`<unk-num>`	has a digit
`<unk-ion>`	ends in *ion*
`<unk-ic>`	ends in *ic*
`<unk-ly>`	ends in *ly*
`<unk-eding>`	ends in *ed* or *ing*
`<unk-allCaps>`	all capitalized
`<unk-firstCaps>`	first letter is capitalized
`<unk-s>`	ends in *s*
`<unk-specChar>`	contains non-alphanumeric character
`<unk-ion>`	catch-all class for remaining unknown tokens

Figure 3: Clusters for OOV in clustering strategy

Results: Clustering OOV strategy

Model	TH 6 Accuracy	TH 3 Accuracy
2-gram Katz	92.44%	93.18%
2-gram Kneser Ney	92.62%	93.33%
2-gram Witten Bell	92.61%	93.32%
3-gram Katz	93.21%	93.85%
3-gram Kneser Ney	93.36%	94.00%
3-gram Witten Bell	93.33%	93.96%
4-gram Katz	92.72%	93.42%
4-gram Kneser Ney	93.00%	93.68%
4-gram Witten Bell	92.85%	93.54%

Figure 4: Test accuracy using clustering OOV strategy

Improvements

- Clearly we are not close to state of the art test performance
- OOV frequency threshold needs to be set based on performance
- OOV clustering should also be done so
- Pruning of n-gram rescoring models
- Including additional information like word lemmas

Questions?

References

- W. Nelson Francis, Henry Kucera *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers.* Brown University. 1979.
- Robert Bley-Vroman *http://www.sls.hawaii.edu/bley-vroman/brown_corpus.html*
- Daniel Jurafsky, James H. Martin *Speech and Language Processing: An Introduction to Natural Language processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall. 2009.
- Evelyne Tzoukermann, Dragomir Radev. *Use of Weighted Finite State Transducers in Part of Speech Tagging*. Cambridge University Press. February 1997.
- André Kempe *Part-of-Speech Tagging with Two Sequential Transducers* Xerox Research Centre Europe. 2000.

References


Miikka Silfverberg, Krister Lindén *Part-of-Speech Tagging using Parallel Weighted Finite-State Transducers*. University of Helsinki. 2010.


Mehryar Mohri, Fernando Pereira, Michael Riley. *Weighted Automata in Text and Speech Processing*. AT&T Research. 1996.


Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut and Mehryar Morhi. *OpenFST: A General and Efficient Weighted Finite-State Transducer LIbrary* Proceedings of the Ninth International Conference on Implementation and Application of Automata, (CIAA 2007), volume 4783 of Lecture Notes in Computer Science, pages 11-23. Springer, 2007. `http://www.openfst.org`

References


Brian Roark, Richard Sproat, Cyril Allauzen, Michael Riley, Jeffrey Sorensen and Terry Tai. 2012. *The OpenGRM open0source finite-state grammar software libraries*. In Proceedings of the ACL 2012 System Demonstrations, pp. 61-66. `http://www.opengrm.org`


Eric Brill. *A Simple Rule-Based Part of Speech Tagger*. In ANLC'92 Proceedings of the third conference on Applied natural language processing. University of Pennsylvania. 1992.


Mehryar Mohri, Fernando Pereira, Michael Riley. *Speech Recognition with Weighted Finite-State Transducers* Spring Handbook on Speech Processing and Speech Communication. 2008.