Tim Macdonald, Leanne Miller, and Kate Henson
Dr. VanDrunen
CS 394: Computational Linguistics
4 December, 2013

Project Three: Using HMMs to POS-tag

To train our POS tagger, we used a corpus consisting of three Oz novels written by Frank L. Baum which was then tagged automatically by the Natural Language Toolkit (NLTK). This corpus was easily available, and seemed to be a reasonable sample of written English.

We did not use the full Penn-Treebank set of tags (with which the corpus was tagged), instead using the following reduced set:

Our Tag	Penn-Treebank Tag(s)	Description
N	NN, NNS, NNP, NNPS, PRP, EX, WP, FW, UH	Noun-like
V	VB, VBD, VBG, VBN, VBP, VBZ, MD, TO	Verb-like
AJ	POS, PRPS, WPS, JJ, JJR, JJS, DT, CD, PDT, WDT, LS	Adjective-like
AV	RB, RBR, RBS, WRB	Adverb-like
G	‘’, (,), ,, :, ``, “, SYM, $, #	Symbol (but not E)
E	., !, ?	End of sentence
P	IN, RP	Preposition/particle
C	CC	Conjunction

This is based on the reduced set used by VanDrunen (which was itself motivated by a desire to simplify the HMM), but with a few key modifications, the most notable being the addition of the P and C tags and the deletion of the S (symbol) and I (interjection) tags. We did not find the G/S distinction meaningful, our corpus was lacking in interjections, and conjunctions and prepositions seemed sufficiently different from adverbs to merit distinct categories.

We employed additive smoothing for calculating probabilities in the HMM, with k = 1. This appears on lines 125 and 137 of pos.py, and resembles:

```
defaultdict(lambda: K / (totals[tag] + K * len(word_tag_tally[tag].values())))
```

Or in more standard notation:

$$\frac{0 + k}{T + k\theta}$$

Where T is the number of tokens present in the corpus of a given tag and θ is the number of types present in the corpus of a given tag.

We tested this against a small hand-tagged text (137 tokens). Our tagger achieved scores of 70.8%, 69.3%, and 70.1% accuracy when trained with a quarter, half, and whole corpus. The NLTK achieved 83.9% accuracy. It is surprising that the accuracy decreases with larger corpus sizes, but the magnitude is negligible. A more sensible conclusion is that the difference in corpus size was insufficiently large to be significant.

At first glance this appears to be severely less than the NLTK's accuracy. However, our corpus was based on the NLTK's results. If we assume that the NLTK's tagging of the corpus was also 83.9% accurate, then the score of our tagger is in fact 70.8/83.9, which is 84.4%, which is comparable to the NLTK. This is a superb result. Real accuracy could obviously be increased by training with a more reliably-tagged corpus. Relative accuracy is harder to improve, but small gains may be had with a larger corpus, especially one with more types.