

Computational Linguistics Final

Jean Mark Gawron

May 7, 2009

1 PCFGs and PCFG Parsing

Visit:

[http://www-rohan.sdsu.edu/~gawron/compling/course_core/
...assignments/final09/prob_parsing_assignment.htm](http://www-rohan.sdsu.edu/~gawron/compling/course_core/...assignments/final09/prob_parsing_assignment.htm)

Do the probabilistic context free grammar parsing problems there.

2 HMM Taggers

- 2.1 Do an error analysis of your tagger. Build a confusion matrix (Section 5.7.1, slide 54 of 56 in slp05.pdf) and investigate the most frequent errors. One approach to implementing the program that builds the confusion matrix is to modify the evaluator.py program you were given for the HMM tagging assignment.
- 2.2 Report your results
 - (a) Send me a printout of the confusion matrix for the top ten confusion pairs .
 - (b) Send me 5 examples (if there are 5) of each of the top 10 errors types. Note in particular, VBD vs. VBN may come up; this is illustrated here.
 - (a) John has always_RB liked_VBN beans
 - (b) John always_RB liked_VBD beans
- 2.3 Do a separate analysis of how many of your errors concern unknown words. Note that the simplest way to collect this information is to modify your tagger so that it keeps and reports counts for unknown

words, since you will always know at tagging time which words aren't in the model. Build a confusion matrix specifically for unknown words.

2.4 Report on your unknown words problem. Print and send the confusion matrix for the top 10 confusion pairs for unknown words.

2.5 Come up with 3 ways your tagger might be improved.

- (a) One improvement should involve a different kind of smoothing from the kind we are using. Hint: You will receive little credit if your alternative smoothing idea is “Use add1 smoothing instead of add 0.5 smoothing”). You *must give the new smoothed probability equations for both the word word and word tag models*.
- (b) One improvement should deal specifically with unknown words. For ideas on what to do with unknown words, see Section 5.8.2. You don't have to implement something but your idea should be specific enough to be implementable. Don't say “Use orthographic and morphological information.” Instead, tell me what algorithms/programs you are going to apply to use that information.
- (c) The third idea is up to you. Again, be specific enough so that the next step, implementation, can be taken.

Finally, you should repeat the error analysis and the unknown words error analysis described above on the Wall Street Journal corpus that will be posted on the blackboard site.

You should then explain the difference in your scores on the two datasets (the dataset you used in the HMM tagger programming assignment was the BNC). In particular, address the following questions: (a) Is there a difference in difficulty between the two tagsets? (b) Is there a difference in difficulty in the two datasets?