Assignment 3

L715: Data manipulation for parser improvement

Due: Wednesday, November 9, 2011

I want you, working in groups of 2, to use the Berkeley parser (`http://code.google.com/p/berkeleyparser/`). Think of your assignment as: becoming an expert on using the Berkeley parser.

Here are some websites with tips for using it:

- `http://sandersn.com/blog//index.php/2010/02/11/tips-for-using-the-berkeley-parser`
- `http://www.coli.uni-saarland.de/~yzhang/rapt-ws1011/tutorial1.pdf`

1. Start by unpacking the jar file & exploring. This requires you to:
   (a) Download the appropriate files:
      - the most recent version (java 1.6)
      - the English grammar
      - README.txt
   (b) Unzip berkeleyParser.jar
   (c) Get it working, specifically making sure you can:
      - Parse - try a baby test sentence first.
      - Train - use small sets of files first.
   (d) Explore :)
2. Provide a (1-2-page?) description of how the code corresponds to the split-merge description provided in the Petrov et al paper. (You may also want to look at the follow-up paper, mentioned in the README.txt file.)
   (a) In other words, you need to figure out what parts of the code do what and how they work (together).
3. Train a parser on a treebank with a few different split-merge cycles and report the experiments and results. If possible, you should test the parsers on both in-domain and out-of-domain data. The different results will be useful for the next part of the assignment ...

- Although not required, I encourage you to alter the code in whatever other ways you feel compelled to. e.g., try (slightly) different unknown word handling, try different weightings of parameters, etc.

4. Get a textual version of your different grammar files (see README.txt for how to do that) and describe properties of the grammars.

   - Quantitatively, look at how many rules there are of different types of categories, how many splits there are, how this differs across settings.
   - Qualitatively, figure out why the rules are the way they are. Are there rules which appear to be overfitting the data? Other rules which could stand to be split again?
   - If time, try parsing with some rules removed, added, merged (by hand). What is the effect? Why?