

Assignment 1

L715: Data manipulation for parser improvement

Due: Wednesday, September 21, 2011

1. Read the papers for your topic and start to determine which ones are the most essential. Provide an ordered list of the most important papers and write up a paragraph as to what the essence of the topic is and how these papers are most relevant—most important, best coverage, etc.
2. Go over the “practical material” in the course slides on the basics of parsing and do the following:
 - (a) Find one constituency treebank and one trainable constituency parser. Using whatever data split you want, train and test the parser. Describe what you did and any difficulties encountered.
 - (b) Find one dependency treebank and one trainable dependency parser. Using whatever data split you want, train and test the parser. Describe what you did and any difficulties encountered.

Some useful places to look for data (in addition to the slides):

- On `jones`, there are various treebanks (see <http://jones.ling.indiana.edu/wiki/JonesCorpora>), including these:
 - `/Volumes/Data/Corpora/en/penntreebankv3/parsed/mrg/` (constituency)
 - `/Volumes/Data/Corpora/en/penntreebankv3/dep/` (dependency)
- CoNLL-X free dependency data: http://ilk.uvt.nl/conll/free_data.html