HW#2

Ling571
Deep Processing Techniques for NLP
January 13, 2016

Goals

- Begin development of CKY parser

- First stage: Conversion to Chomsky Normal Form

  - Develop representation for CFG

  - Manipulate/transform grammars

  - Investigate weakly equivalent grammars

Task

- Conversion:
  - Read in grammar rules from arbitrary CFG

  - Convert to CNF

  - Write out new grammar

- Validation:
  - Parse test sentences with original CFG
  - Parse test sentences with CFG in CNF

Approach

- May use any programming language you like
  - As long as it runs on the cluster

- May use existing models to represent rules
  - Need RULE, RHS, LHS, etc
  - NLTK, Stanford

- Conversion code *must* be your own

Data

- ATIS (Air Travel Information System) data
  - Grammar provided in nltk-data
    - Terminals in double-quotes
      - the -> "the"
  - Sentences linked from homework webpage

- NOTE: Grammar is fairly large (193K)
  - Develop on smaller subset !

  - Grammar is also fairly ambiguous
    - Test sentences can have 100 parses

NLTK grammars

- Gr1 = nltk.data.load("grammars/large_grammars/atis.cfg")

- Gr1.productions()[0]
  - ABBCL_NP -> QUANP_DTI QUANP_DTI QUANP_CD AJP_JJ NOUN_NP PRPRTCL_VBG

- Gr1.productions()[0].lhs()
  - ABBCL_NP

- Gr1.productions(lhs=Gr1.productions()[1].lhs())
  - [ADJ_ABL -> only, ADJ_ABL -> such]