Midterm Exam Notes

Coverage: Linguistic Knowledge / Representations & Algorithms, e.g.,

- Morphology / Transducers
- Language Modeling / N-Grams
- Part of Speech / Tagsets & Tagging
- Constituency/ (P)CFGs & Parsing
- Vector Semantics

Types of questions:

True/False

- The Penn Treebank part of speech tags is the only tagset for English.
- Spelling rules can be implemented as finite state transducers.

Short Answer

- Explain and compare smoothing and backoff.
- Why do we usually make a Markov assumption and deal with N-grams?
- What do people use the Penn Treebank for? What are its limitations?

Problem Solving

- Consider the following finite state transducer from the book: (figure)
  - What kind of knowledge of language is this FST representing?
  - Use this FST to parse the four types of input that get you to an accept state. For each of the four parses, show the input and the output, as well as the states that you pass through.

- Consider the following probabilistic context-free grammar (PCFG): (figure)
  - Convert to CNF
  - Show all possible parses of the following sentence.
  - Compute the probability of each of the trees (you can just write an equation).
  - Give one example motivating why and showing how you might want to do parent annotation.