Language and Statistics II: (More) Empirical Methods in Natural Language Processing 11-762

Noah Smith
School of Computer Science
Carnegie Mellon University

Fall 2006

Summary

This course will cover modern empirical methods in natural language processing. It is designed for language technologies students who want to understand statistical methodology in the language domain, and for machine learning students who want to know about current problems and solutions in text processing.

Students will, upon completion, understand how statistical modeling and learning can be applied to text, be able to develop and apply new statistical models for problems in their own research, and be able to critically read papers from the major related conferences (EMNLP and $*$ACL). A recurring theme will be the tradeoffs between computational cost, mathematical elegance, and applicability to real problems. The course will be organized around methods, with concrete tasks introduced throughout.

Target

The course is designed for SCS graduate students. Prerequisite: Language and Statistics (11-761) or permission of the instructor. Recommended: Algorithms for Natural Language Processing (11-711), Machine Learning (15-681, 15-781, or 11-746).

Evaluation

Students will be evaluated in four ways.

Literature review (35%). Each student will individually complete a literature review on a problem within natural language processing that interests him/her. The literature review is expected to be comprehensive and include a problem definition, evaluation, a discussion of available datasets, and a thorough, coherent discussion of existing techniques. Insofar as possible, comparison should be given among different techniques. Current obstacles should be discussed, ideally with insights on tackling or avoiding them. Implementation is not required for this literature review. Suggested topics include: question answering, textual entailment, paraphrase, morphology induction and modeling, syntax-based machine translation, data-oriented models (DOP), syntax-based language modeling. Each student will take on a different topic.

To encourage effective written communication, there will be two rounds of evaluation. A draft will be due about four weeks before the end of the term, so that feedback from the instructor can improve the quality of the final document.

Oral presentation and discussion (25%). Each student will give a $\sim$20-minute oral presentation on his/her literature review. A period of discussion will follow, in which we will aim to find connections between student topics. The driving questions will be: What can be borrowed from one area and applied to another? And what challenges are not being met by current methods?

Assignments (20%). Small projects with a programming component will be assigned about every three weeks.Typically some or all of the software will be available, and students will be expected to run experiments or extend implementation. To encourage creative exploration, some projects may be graded competitively.

Final exam (20%). A final written exam will be given to test basic competence with the technical material covered in the lectures.

Topics

An outline follows. The papers listed will not necessarily be assigned as readings. They are shorthand for approximate material that will be covered under each section of the course, and they are subject to change! Numbers in parentheses are the number of lectures (I assume a twice-weekly meeting of around 75 minutes, and a 14-week semester). Starred items will be postponed or canceled if the schedule lags.

1. Philosophy: the empirical way of thinking about language. (1)
   Abney (1996).

2. Stochastic models for sequences: Markov chains and hidden Markov models. Smoothing. (2)
   Language modeling, part-of-speech tagging, shallow parsing, named entity recognition.

3. Log-linear models, conditional random fields. Regularization. Numerical optimization. (3)
   Chen & Rosenfeld (2000), Lafferty et al. (2001), Sha & Pereira (2003), etc.

4. Weighted finite-state machines and transducers. (2)
   Mohri (1997), Eisner (2002).

5. Stochastic and weighted context-free grammars (and beyond). (3)
   Phrase structure parsing, dependency parsing, Charniak (1997), Collins (1999), Eisner (2002), Abney (1997).

6. Weighted dynamic programming. (2)
   Goodman (1999), Caraballo & Charniak (1998), Klein & Manning (2003), Eisner et al. (2005).

7. Discriminative training and reranking. (2)
   Collins (2000), Charniak & Johnson (2005), Taskar et al. (2004), McDonald et al. (2005).

8. *Transformation-based learning, local classifiers. (1)
   Brill (1992), Punyakanok et al. (2005).

9. Expectation-maximization: learning models from unannotated examples. (2)
   Merialdo (1994), Carroll & Charniak (1992), translation models: Brown et al. (1993), Melamed (2000).

10. *Other unsupervised methods: contrastive estimation, deterministic annealing, structural annealing, state merging and splitting. (3)
    Smith & Eisner (2005, 2006), Chen (1995), Stolcke and Omohundro (1994). Learning phrasal translations.

11. Bootstrapping, cotraining, and semi-supervised learning. (2)
    Yarowsky (1995), cross-lingual learning.

The final lectures of the course will be devoted to the oral presentations and discussion.

Readings

Manning and Schütze's *Foundations of Statistical Natural Language Processing* will be recommended for background reading during parts of the course, though many of the techniques taught are predated by that book. Readings will be suggested from recent conferences and journal articles, perhaps also chapters from Jurafsky & Martin's *Speech and Language Processing*, MacKay's *Information Theory, Inference, and Learning Algorithms*, Klavans and Resnik's *The Balancing Act*, or other texts. No particular readings will be mandatory, though a great deal of reading will be required for completion of the literature review.