Linguistics 696f
Hammond
Spring '03

Statistical Natural Language Processing

Course description

This course introduces the key concepts underlying statistical natural language processing. These include, e.g. n-gram models, smoothing, Hidden Markov models, and higher-order language models. Our goals are twofold. On the one hand, we want to be able to use these modeling techniques for concrete goals. Second, we want to understand the potential theoretical consequences of these techniques: how do they fare as theories of language? (No mathematical background beyond high school algebra is needed.)

instructor	Mike Hammond
office	Douglass 204
hours	Wednesday 9:00-10:00 or by appointment
email	hammond@u.arizona.edu
class website	http://linguistics.arizona.edu/~hammond/ling696f-sp03/

Requirements

This is new material for most of us and a novel setting for all of us. The requirement structure is designed to be attentive to these factors.

assignment	value	due date
small exercise #1	5%	Feb 24
small exercise #2	5%	Mar 24
project prospectus	10%	Apr 14
presentation	10%	May 5
final project	70%	May 12

The final project can be a traditional final paper, but can also be a programming or experimental project of analogous scope. The prospectus is a one-page outline of the final project.

Schedule

Except for the first week, readings should be done *before* the week they are listed for. All readings except the text are available on the class website.

Week	*Date*	*Topic*	*Readings*	*Due*
1	1/20	**MLK day: holiday**		
2	1/27	Introduction & automata	[Abn96], [Cha93, ch.1]	
3	2/3	Probability theory	[Cha93, ch.2]	
4	2/10	N-gram models	[Cha93, ch.3]	
5	2/17	N-grams continued		
6	2/24	Smoothing	[CG98]	Exercise #1
7	3/3	Toolkits	[Sto02], [CR97]	
8	3/10	HMMs		
	3/17	**Spring break**		
9	3/24	HMM algorithms	[Cha93, ch.4]	Exercise #2
10	3/31	PCFGs	[Cha93, ch.5]	
11	4/7	PCFG algorithms	[Cha93, ch.6]	
12	4/14	Applications	[Cha93, ch.8,ch.9]	Prospectus
13	4/21	Are these theories?	TBA	
14	4/28	My world	[CP97], [FLP00]	
15	5/5	Presentations		presentations
16	5/12	**No class**		final projects

Readings

[Abn96] Steven Abney. Statistical methods and linguistics. In Judith Klavans and Philip Resnik, editors, *The Balancing Act*, pages 1–26. MIT Press, Cambridge, 1996.

[CG98] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. Technical report, Harvard University, 1998.

[Cha93] Eugene Charniak. *Statistical Language Learning*. MIT Press, Cambridge, 1993.

[CP97] J. Coleman and J. Pierrehumbert. Stochastic phonological grammars and acceptability. In *Computational phonology: Third meeting of the ACL special interest group in computational phonology*, pages 49–56. Association for Computational Linguistics, Somerset, 1997.

[CR97] P.R. Clarkson and R. Rosenfeld. Statistical language modeling using the CMU-Cambridge toolkit. In *Proceedings ESCA Eurospeech*, 1997.

[FLP00] S. Frisch, N.R. Large, and D.B. Pisoni. Perception of wordlikeness: effects of segment probability and length on the processing of nonwords. *Journal of Memory and Langauge*, 42:481–496, 2000.

[Sto02] Andreas Stolcke. SRILM - an extensible language modeling toolkit. In *Proc. Intl. Conf. Spoken Language Processing*, Denver, 2002.