Winter 2010

CS 771-01: Natural Language Processing Techniques

Shaojun Wang
Wright State University - Main Campus, shaojun.wang@wright.edu

CS771: NATURAL LANGUAGE PROCESSING TECHNIQUES
WINTER 2010

INFORMATION
SYLLABUS
ASSIGNMENTS
PROJECT

TENTATIVE SYLLABUS

Day	Topic	Reading	Optional Reading
1/04/10	Introduction, basics of information theory	JM: 1; MS: 1-3	Statistical methods in Encyclopedia of Cognitive Science by Steve Abney.
1/06/10	Source-channel model, zipf's law and n-grams	JM: 9.1, 25.3, 4.4, 4.10-11; MS: 2.2.4, 1.4.3	Information retrieval as statistical translation Adam Berger and John Lafferty, SIGIR 1999
1/11/10	Smoothing techniques: Laplace, Good-Turing	JM: 4.5; MS: 6.2	Good, Jelinek, Mercer, and Robins on Turing's estimate of probabilities in American Journal of Mathematical and Management Sciences, 11, 229-308, 1991 by Arthur Nadas; Always Good Turing: asymptotically optimal probability estimation in Science. 302(5644):427-431 by Alon Orlitsky et al.
1/13/10	Deleted linear interpolation; EM algorithm for linear interpolation	JM: 4.6	Interpolated Estimation of Markov Source Parameters from Sparse Data in Pattern Recognition in Practice, 381-397, 1981 by F. Jelinek and R. Mercer; Convexity, maximum likelihood and all that By Adam Berger
1/18/10	Martin Luther King Day. University closed		
1/20/10	Back-off: Katz, Kneser-Ney; Limitations of n-grams	JM: 4.7-9; MS: 6.3	A hierarchical Bayesian language model based on Pitman-Yor processes by Yee Whye Teh, ACL 2006; An empirical study of smoothing techniques for language modeling in Computer Speech & Language. 13(4):319-358. 1999 by S. Chen and J.

			Goodman; <u>Up from trigrams! - the struggle for improved language models</u>. EUROSPEECH, 1037-1040, 1991 by Fred Jelinek
1/25/10	EM algorithm; Hidden Markov models	JM: 6; MS: 9	
1/27/10	Part-of-speech tagging, Named Entity recognition; Conditional random fields	JM: 6.22	<u>Conditional random fields: probabilistic models for sand labeling sequence data</u> by Lafferty et al., ICML 2001; <u>Identifying gene and protein mentions in text using conditional random fields</u> in Bioinformatics, 2005 by McDonald and Pereira
2/01/10	Semi-supervised conditional random fields	JM: 6	<u>Semi-supervised conditional random fields for improved sequence segmentation and labeling</u> by Jiao et al., ACL 2006; <u>Efficient computation of entropy gradient for semi-supervised conditional random fields</u> in NAACL/HLT, 2007 by Mann and McCallum
2/03/10	Grammars and parsing	JM: 12-13	
2/08/10	Probabilistic context-free grammars: inside-outside algorithm	JM: 14.1-5; MS: 11	Lafferty's notes
2/10/10	Probabilistic lexicalized CFGs	JM: 14.6	<u>Head-driven statistics methods for natural language parsing</u> in Computational Linguistics, 29(4): 589-637, 2003 by Collins; <u>Intricacies of Collins' parsing model</u> in Computational Linguistics, 30(4): 479-511, 2004 by Bikel
2/15/10	Parser based language modeling: structured language model, generalized inside-outside algorithm		<u>Structured language model</u> in Computer Speech and Language, 14(4):283-332, 2000 by Chelba and Jelinek; <u>Stochastic analysis of structured language modeling</u> in Mathematical Foundations of Speech and Language Processing, 37-72, 2004 by Jelinek
2/17/10	Word sense disambiguation	JM: 20.1-5	<u>Unsupervised word sense disambiguation rivaling supervised methods</u> in ACL 1995 by Yarowsky
2/22/10	Topic models: LSA;	MS: 15.4	<u>Indexing by latent semantic</u>

	Topic based language modeling		analysis in Journal of the American Society for Information Science, 41(6):391-407, 1990 by Deerwester et al.; Exploiting latent semantic information in statistical language modeling in Proceedings of the IEEE, 88 (8):1279-1296, 2000 by Bellegarda
2/24/10	Topic models: PLSA; Composite language model		Unsupervised learning by probabilistic latent semantic analysis in Machine Learning. 42 (1):177-196, 2001 by Hoffman
3/01/10	Topic models: LDA		Latent Dirichlet allocation in Journal of Machine Learning Research, 3:993-1022, 2003 by Blei et al.
3/03/10	Machine translation: IBM models 1-5; phrase- and parsing based models	JM: 25.5, 25.11	The mathematics of statistical machine translation: parameter estimation in Computational Linguistics, 29(2): 263-311, 1993 by Brown et al.; Hierarchical phrase-based translation in Computational Linguistics, 33 (2):201-228, 2007 by Chiang; Statistical machine translation at WikiPedia
3/08/10	Machine translation: decoding algorithms	JM 25.8	
3/10/10	Machine translation: BLEU	JM 25.9	BLEU: a Method for Automatic Evaluation of Machine Translation in ACL 2002 by Papineni et al.; BLEU at WikiPedia
3/15/10	Project presentations		