Grammar Induction

General Information

Organizer: Bryan Jurish (moocow@ling.uni-potsdam.de)
Office: II.24.180
Telephone: (0331) 977 2180
Office Hours: by appointment
Course Web Site:
http://www.ling.uni-potsdam.de/~moocow/class/grammar-induction

Course Outline

Conventional syntactic theories assume a structure inherent in natural language which is innate, i.e. known *a priori* by a human learner. Much work in computational linguistics has thus dealt with the formal characterization and implementation of such hypothesized innate linguistic knowledge. In this course, we will focus instead on *a posteriori* ("unsupervised") approaches to language learning which assume minimal (if any) prior knowledge on the part of the learner.

After a brief review of the basics of probability and information theory, we will discuss several alternative theoretical learning paradigms and their relevance to natural language induction. The second phase of the course will be concerned with the induction of lexical (syntactic) categories by means of *clustering*. The remainder of the course will focus on the unsupervised acquisition of *language models* – in particular *stochastic finite state automata* and *stochastic context-free grammars* – by means of *expectation-maximization*, *genetic algorithms*, *alignment-based learning*, and additional linguistically-oriented algorithms.

Course Requirements

- In-class presentation **and** written paper are required for acquisition of a *Schein*.
- Topics for in-class presentations, papers, and programming projects must be arranged with me.
- Students presenting a topic in class must meet with me to discuss the respective topic at least one week before the presentation is scheduled.

Course Syllabus

Week	Date	Topic(s)	Reading
1	14.04.	Administrivia	
2	21.04.	Probability Theory	Manning and Schütze (1999, Ch. 1,2.1)
3	28.04.	Information Theory	Manning and Schütze (1999, Ch. 2.2)
4	05.05.	*no class*	
5	12.05.	Learnability	Gold (1967); Angluin and Smith (1983); Valiant (1984)
6	19.05.	Lexical Categories 1	Brown et al. (1992)
7	26.05.	Lexical Categories 2	Finch and Chater (1993); Roberts (2002)
8	02.06.	RLs: EM	Manning and Schütze (1999, Ch. 9); Vidal et al. (2004a,b)
9	09.06.	RLs: ALGERIA, RLIPS	Carrasco and Oncina (1994, 1999)
10	16.06.	RLs: MDI, MALGERIA	Thollard et al. (2000); Habrard et al. (2003)
11	23.06.	CFLs: EM	Manning and Schütze (1999, Ch. 11); Charniak (1993, Ch. 6,7)
12	30.06.	CFLs: GAs	Lankhorst (1994)
13	07.07.	CFLs: ABL	van Zaanen (2000)
14	14.07.	CFLs: DGs	Klein and Manning (2004)

References

D. Angluin and C. H. Smith. Inductive inference: theory and methods. *ACM Computing Surveys*, 15(3):237–269, September 1983.

P. F. Brown, V. J. D. Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4): 467–479, 1992.

R. Carrasco and J. Oncina. Learning stochastic regular grammars by means of a state merging method. In *Proc. 2nd International Colloquium on Grammatical Inference - ICGI '94*, volume 862, pages 139–150. Springer-Verlag, 1994.

R. C. Carrasco and J. Oncina. Learning deterministic regular grammars from stochastic samples in polynomial time. *RAIRO (Theoretical Informatics and Applications)*, 33(1):1–20, 1999.

E. Charniak. *Statistical Language Learning*. MIT Press, Cambridge, MA, 1993.

S. Finch and N. Chater. Learning syntactic categories: a statistical approach. In *Neurodynamics and Psychology*, pages 295–322. Harcourt Brace, London, 1993.

E. M. Gold. Language identification in the limit. *Information and Control*, 10: pp. 447–474, 1967.

A. Habrard, M. Bernard, and M. Sebban. Improvement of the state merging rule on noisy data in probabilistic grammatical inference. In *ECML*, pages 169–180, 2003.

D. Klein and C. Manning. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting of the ACL*, 2004.

B. Krenn and C. Samuelsson. *The Linguist's Guide to Statistics*. URL `http://www-old.coli.uni-saarland.de/~thorsten/c-alg/stat_cl.ps.gz`. Unpublished manuscript, 1997.

M. Lankhorst. Breeding grammars: Grammatical inference with a genetic algorithm. Technical Report CS-R9401, C.S. Department, Univ. of Gronigen, The Netherlands, 1994.

C. D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, 1999.

T. M. Mitchell. *Machine Learning*. McGraw-Hill, Boston, MA, 1997.

A. Roberts. Automatic acquisition of word classification using distributional analysis of content words with respect to function words. Technical report, School of Computing, University of Leeds, 2002.

F. Thollard, P. Dupont, and C. de la Higuera. Probabilistic DFA inference using Kullback-Leibler divergence and minimality. In *Proc. 17th International Conf. on Machine Learning*, pages 975–982. Morgan Kaufmann, San Francisco, CA, 2000.

L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984. ISSN 0001-0782.

M. van Zaanen. ABL: Alignment-based learning. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING)*, pages 961–967, aug. 2000.

E. Vidal, F. Thollard, C. de la Higuera, , F. Casacuberta, and R. C. Carrasco. Probabilistic finite-state machines – Part I. *IEEE Trans. on Pattern analysis and Machine Intelligence*, to appear, 2004a.

E. Vidal, F. Thollard, C. de la Higuera, , F. Casacuberta, and R. C. Carrasco. Probabilistic finite-state machines – Part II. *IEEE Trans. on Pattern analysis and Machine Intelligence*, to appear, 2004b.