Seminar in Natural Language Processing: Finite-State Technology

Shuly Wintner

Spring, 2001: Tuesday, 16:00–18:00, Rabin 7037
`http://cs.haifa.ac.il/ shuly/teaching/01/seminar/`

Objectives

The seminar will cover classic and contemporary topics in finite-states approaches to natural language processing. We will read papers that deal with the application of finite-state technology in areas such as phonology, morphology and syntax of natural languages. We will also read some papers that discuss the mathematical and computational properties of finite-state devices (automata and transducers).

Schedule

Week 1: introduction I will give a brief introduction to finite-state automata and transducers, and exemplify their uses in natural language processing. I will also touch upon some mathematical and computational aspects of finite-state devices such as closure properties, minimization, determinization etc. The introduction will be based on Roche and Schabes (1997c) and Karttunen (1991).

Week 2 Karttunen et al. (1996). A fairly easy paper that relates regular expressions and relations to finite automata and transducers, and exemplifies their use in several language engineering applications.

Week 3 Koskenniemi (1983). This is the classic presentation of *Two-Level Morphology*, and an exposition of the two-level rule formalism. You will have to deliver chapters 1–2, with examples from chapter 3. The book is rather easy to read, the math is very easy and the linguistics will require some preparation.

Week 4, 5 Kaplan and Kay (1994) – 2 students. A classic work that sets the very basics of finite-state phonology, referring to automata, transducers and two-level rules. Both the mathematics and the linguistics require some sophistication, but the paper is very well written and provides plenty of examples.

Week 6 Karttunen (1997). A paper that introduces the *replace* operator to the calculus of regular expressions. No linguistic background is needed, and the math isn't very hard.

Week 7 Roche and Schabes (1997a). The use of finite-state transducers for a particular application: part of speech tagging. The paper is computationally oriented.

Week 8 Pereira and Riley (1997). The use of weighted finite-state automata for a particular application: speech recognition. The paper is mathematically oriented.

Week 9 Beesley (1996), Beesley (1998a) and Beesley (1998b). Three related papers that deal with finite-state morphological analysis of Arabic. Knowledge of Arabic is recommended.

Week 10 Kiraz (2000). A novel approach to a computational treatment of non-concatenativemorphology with motivation and examples from Semitic languages.

Week 11 Pereira and Wright (1997). A very elegant paper specifying how context-free grammars can be approximated by finite-state machines. Requires some knowledge of parsing theory.

Week 12 Nederhof (2000). A survey of finite-state approximations of context-free grammars. A pretty challenging paper.

Week 13 Daciuk et al. (2000). An algorithm for constructing minimal automata from lists of words, sorted or unsorted. Computationally oriented and very clear.

Week 14 Discussion and future ideas.

Additional papers

Mohri (1996), Mohri (1997a) and Mohri (1997b) – 2–3 students. Three related papers that present some mathematical and computational properties of sequential and subsequential transducers and exemplify their uses in natural language processing. Very little linguistics is required, but the mathematics can get hairy.

Administration

Attendance is compulsory and absence will be penalized.
Grading will be based on comprehension of the assigned article, ability to explain it and quality of the presentation. Efforts to understand other papers will be generously rewarded.
Presentations can be either in Hebrew or in English, but must be coherent. Use a language you have full command of!
Computerized presentations are allowed but not required.
Make sure you understand the paper before you start planning your presentation.
A summary of the paper is not required but might help you tremendously.

References

Beesley, Ken. 1996. Arabic finite-state morphological analysis and generation. In *Proceedings of COLING-96, the 16th International Conference on Computational Linguistics*, Copenhagen.

Beesley, Ken. 1998a. Arabic morphological analysis on the internet. In *Proceedings of the 6th International Conference and Exhibition on Multi-lingual Computing*, Cambridge, April.

Beesley, Kenneth R. 1998b. Arabic morphology using only finite-state operations. In Michael Rosner, editor, *Proceedings of the Workshop on Computational Approaches to Semitic languages*, pages 50–57, Montreal, Quebec, August. COLING-ACL'98.

Daciuk, Jan, Stoyan Mihov, Bruce W. Watson, and Richard E. Watson. 2000. Incremental construction of minimal acyclic finite-state automata. *Computational Linguistics*, 26(1):3–16, March.

Kaplan, Ronald M. and Martin Kay. 1994. Regular models of phonological rule systems. *Computational Linguistics*, 20(3):331–378, September.

Karttunen, Lauri. 1991. Finite-state constraints. In *Proceedings of the International Conference on Current Issues in Computational Linguistics*, Universiti Sains Malaysia, Penang, Malaysia, June. Available from `http://www.xrce.xerox.com/research/mltt/fst/fsrefs.html`.

Karttunen, Lauri. 1997. The replace operator. In *Finite-State Language Processing* (Roche and Schabes, 1997b), chapter 4, pages 117–147.

Karttunen, Lauri, Jean-Pierre Chanod, Gregory Grefenstette, and Anne Schiller. 1996. Regular expressions for language engineering. *Natural Language Engineering*, 2(4):305–328.

Kiraz, George Anton. 2000. Multitiered nonlinear morphology using multitape finite automata: a case study on Syriac and Arabic. *Computational Linguistics*, 26(1):77–105, March.

Koskenniemi, Kimmo. 1983. *Two-Level Morphology: a general computational model for word-form recognition and production*. The department of general linguistics, University of Helsinki.

Mohri, Mehryar. 1996. On some applications of finite-state automata theory to natural language processing. *Natural Language Engineering*, 2(1):61–80.

Mohri, Mehryar. 1997a. Finite-state transducers in language and speech processing. *Computational Linguistics*, 23(2):269–312, June.

Mohri, Mehryar. 1997b. On the use of sequential transducers in natural language processing. In *Finite-State Language Processing* (Roche and Schabes, 1997b), chapter 12, pages 355–381.

Nederhof, Mark-Jan. 2000. Practical experiments with regular approximation of context-free languages. *Computational Linguistics*, 26(1):17–44, March.

Pereira, Fernando C. N. and Michael D. Riley. 1997. Speech recognition by composition of weighted finite automata. In *Finite-State Language Processing* (Roche and Schabes, 1997b), chapter 15, pages 431–453.

Pereira, Fernando C. N. and Rebecca N. Wright. 1997. Finite-state approximation of phrase-structure grammars. In *Finite-State Language Processing* (Roche and Schabes, 1997b), chapter 5, pages 149–174.

Roche, Emmanuel and Yves Schabes. 1997a. Deterministic part-of-speech tagging with finite-state transducers. In *Finite-State Language Processing* (Roche and Schabes, 1997b), chapter 7, pages 205–240.

Roche, Emmanuel and Yves Schabes, editors. 1997b. *Finite-State Language Processing*. Language, Speech and Communication. MIT Press, Cambridge, MA.

Roche, Emmanuel and Yves Schabes. 1997c. Introduction. In *Finite-State Language Processing* (Roche and Schabes, 1997b), chapter 1, pages 1–65.