Introduction to the Special Issue on Finite-State Methods in NLP

Lauri Karttunen*
Xerox Research Centre Europe

Kemal Oflazer†
Bilkent University

The idea for this special issue came up during the preparations of the International Workshop on Finite-State Methods in Natural Language Processing, that was held at Bilkent University in Ankara, Turkey in the summer of 1998. The number of the submissions had exceeded our initial expectations and we were able to select quite a good set of papers from those submitted. Further, the workshop and the preceding tutorial by Kenneth Beesley, on finite-state methods, was attended by quite a large number of participants. This led us to believe that interest in the theory and applications of finite-state machinery was alive and well, and that some of the papers from this workshop along with further additional submissions could make a very good special issue for this journal. The five papers in this issue are the result of this process.

The last decade has seen a quite a substantial surge in the use of finite-state methods in all aspects of natural language applications. Fueled by the theoretical contributions of Kaplan and Kay (1994), Mohri's recent contributions on the use of finite-state techniques in various NLP problems (Mohri 1996, 1997), the success of finite-state approaches especially in computational morphology, for example, Koskenniemi (1983), Karttunen (1983), and Karttunen, Kaplan, and Zaenen (1992), and, finally, the availability of state-of-the-art tools for building and manipulating large-scale finite-state systems (Karttunen 1993; Karttunen and Beesley 1992; Karttunen et al. 1996; Mohri, Pereira, and Riley 1998; van Noord 1999), recent years have seen many successful applications of finite-state approaches in tagging, spell checking, information extraction, parsing, speech recognition, and text-to-speech applications. This is a remarkable comeback considering that in the dawn of modern linguistics (Chomsky 1957), finite-state grammars were dismissed as fundamentally inadequate. As a result, most of the work in computational linguistics in the past few decades has been focused on far more powerful formalisms.

Recent publications on finite-state technology include two collections of papers (Roche and Schabes 1997; Kornai 1999) with contributions covering a wide range of these topics. This special issue, we hope, will add to these contributions.

The five papers in this collection cover many aspects of finite-state theory and applications. The papers *Treatment of Epsilon Moves in Subset Construction* by van Noord and *Incremental Construction of Minimal Acyclic Finite-State Automata and Transducers* by Daciuk, Watson, Watson, and Mihov, address two fundamental aspects in the construction of finite-state recognizers. Van Noord presents results for various methods for producing a deterministic automaton with no epsilon transitions from a nondeterministic automaton with a large number of epsilon transitions, especially those resulting from finite-state approximations of context-free and more powerful formalisms. Daciuk et al. present a new method for constructing minimal, deterministic, acyclic

* 6, chemin de Maupertuis, 38240, Meylan, France
† Bilkent, TR-06533, Ankara, Turkey

finite-state machines from a list of input strings, in a single pass. *Practical Experiments with Regular Approximations of Context-free Languages* by Nederhof, presents evaluations of various regular approximation algorithms on actual grammars, providing insights into pros and cons of such algorithms. *Multitiered Nonlinear Morphology Using Multitape Finite Automata: A Case Study on Semitic* by Kiraz presents the formalism and implementation of an approach for dealing with nonlinear phenomena found in the morphology of semitic languages and compares his approach with other systems that have been proposed for the same languages. Finally, *Learning Dependency Translation Models as Collections of Finite-State Head Transducers* by Alshawi, Bangalore, and Douglas, presents an application of the finite-state transducer framework in a machine translation task where weighted finite-state head transducers induced from a corpus of aligned parallel sentences are used to recursively map headwords from the source to the target language.

Our guest editorial board for this issue included Ken Beesley, Eric Brill, Eva Ejerhed, George Kiraz, András Kornai, Mehryar Mohri, Mark-Jan Nederhof, Martin Kay, Ron Kaplan, and Atro Voutilainen; we received additional help from many other reviewers. Julia Hirschberg, editor-in-chief of *Computational Linguistics*, helped us through all aspects of the selection process, guiding us around many intricate issues. We thank the guest editorial board, the additional reviewers, and Julia for their superb contributions. We hope you find this special issue well worth the effort.

References

Chomsky, Noam. 1957. *Syntactic Structures*. Mouton, The Hague.

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

Karttunen, Lauri. 1983. KIMMO: A general morphological processor. *Texas Linguistic Forum*, 22:163–186.

Karttunen, Lauri. 1993. Finite-state lexicon compiler. Technical Report, XEROX Palo Alto Research Center, April. Available at http://www.xrce.xerox.com/research/mltt/fsSoft.

Karttunen, Lauri and Kenneth. R. Beesley. 1992. Two-level rule compiler. Technical Report, XEROX Palo Alto Research Center. Available at http://www.xrce.xerox.com/research/mltt/fsSoft.

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

Karttunen, Lauri, Ronald M. Kaplan, and Annie Zaenen. 1992. Two-level morphology with composition. In *COLING-92: Papers Presented to the 15th* [sic] *International Conference on Computational Linguistics*, volume 1, pages 141–148, Nantes, France. International Committee on Computational Linguistics.

Kornai, András, editor. 1999. *Extented Finite State Models of Language*. Cambridge University Press, Cambridge, England.

Koskenniemi, Kimmo. 1983. Two-level morphology: A general computational model for word form recognition and production. Publication No: 11, 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*, 2:1–20.

Mohri, Mehryar. 1997. Finite-state transducers in language and speech processing. *Computational Linguistics*, 23(2):269–311, June.

Mohri, Mehryar, Fernando C.N. Pereira, and Michael Riley. 1998. A rational design for a weighted finite–state transducer library. In Derick Wood and Sheng Yu, editors, *Automata Implementation*. Lecture Notes in Computer Science, Number 1436. Springer Verlag, pages 144–158.

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

van Noord, Gertjan. 1999. FSA6: Finite state automata utilities (version 6) manual. Available at http://odur.let.rug.nl/vannoord/Fsa/Manual.