Extended finite state models of language

ANDRÁS KORNAI
IBM Almaden Research Center
650 Harry Road, San Jose, CA 95120, USA
kornai@almaden.ibm.com

(*Received December 1 1997; Revised February 9 1997*)

In spite of the wide availability of more powerful (context free, mildly context sensitive, and even Turing-equivalent) formalisms, the bulk of the applied work on language and sublanguage modeling, especially for the purposes of recognition and topic search, is still performed by various finite state methods. In fact, the use of such methods in research labs as well as in applied work actually increased in the past five years. To bring together those developing and using extended finite state methods to text analysis, speech/OCR language modeling, and related CL and NLP tasks with those in AI and CS interested in analyzing and possibly extending the domain of finite state algorithms, a workshop was held in August 1996 in Budapest as part of the European Conference on Artificial Intelligence (ECAI'96).

The present special issue of *Natural Language Engineering*, as well as the companion volume to be published in the ACL Studies in Natural Language Processing series this year, grew out of the proceedings of this workshop, available from the von Neumann Society of Computer Science (Báthori u. 16, H-1054 Budapest, Hungary) in hard copy, or at `www.cs.rice.edu/~andras/ecai.html` on the web. JNLE readers whose interest in the subject of finite state technologies is aroused by this issue are advised to look at these proceedings, since they contain several excellent papers that could not be included here because of space constraints or because the authors felt that their subsequent work took a direction that they no longer consider the workshop paper fully representative of their current thinking. In particular, we call attention to the tutorial paper by Jelinek (excerpted from a his forthcoming book (Jelinek 1977)), the paper by Mohri, Pereira, and Riley describing the AT&T/Bell Labs approach to language modeling using weighted transducers, and the paper by Oehrle on binding and anaphora.

Even without these papers, the sheer size of the proceedings made it impossible to include the same material in this issue of JNLE, and the participants were asked to prepare shorter versions (in some cases, extended abstracts) for inclusion here. A full version of these papers, taking into account the comments received at the workshop, will be published later this year by Cambridge University Press. In addition, a formal Call For Papers yielded several new papers for this issue and for the volume, making the original proceedings, the current issue, and the forthcoming volume independently valuable for researchers in this area.

To understand some of the main trends in finite state NLP it is worth looking back at the origins of the field. Though neither Mealy (1955) nor Kleene (1956) had NL applications in mind, finite state methods were applied in this domain as early as 1958. The rediscovery of this work (see **Joshi**'s paper in this issue and Karttunen's comments in the companion volume) has been one of the pleasant surprises of the ECAI workshop. In the early sixties, however, finite state models were soon submerged in a flood of transformational models. At that time neither careful attendance to linguistic detail nor husbanding of computational resources held much appeal, and the excitement generated by the breathtaking pace of development from Syntactic Structures to Aspects and the Standard Model, the Extended Standard Model, and the Revised Extended Standard Model kept most computational linguists too busy to think through the implications.

It is hard to speculate about such matters, but it is quite conceivable that the finite state approach to NLP would have lost all credibility, were it not for the extraordinary impact of Thompson (1968) and the `grep` family of unix tools. While theoretical linguists accepted the arguments put forth in Miller and Chomsky (1963) at face value, from the seventies it became part of the received computer science wisdom that if you want to do something with text you need to build finite automata. By making his implementation of `regexp(3)` freely redistributable, Spencer (1986) transmitted this wisdom to the free software movement, and subsequent works including GNU `flex` and `agrep` (Wu and Manber 1992) have spread to many corners of computer science from compilers to protocols. In this issue, this trend is represented by the FIRE Lite toolkit described by **Watson**, which finally brings computations involving automata with hundreds of thousands or even millions of states outside the confines of highly proprietary development environments. As automata grow in size, it is becoming increasingly important to develop tools for their testing and debugging, and the work described in **Vilares** *et al.* is a good first step in this direction.

Given the dominant position of finite state technologies in topic search, in retrospect it is hard to understand why mainstream syntactic theory continued to shun finite state methods throughout the seventies and eighties, but in fact these methods reappeared on the scene through a back door left open by the context sensitive rule systems of phonology. Only two years after the seminal Sound Pattern of English (Chomsky and Halle 1968), Johnson (1970) demonstrated that the context sensitive machinery of SPE can be replaced by a much simpler one, based on finite state transducers (FSTs), and independently the same conclusion was reached by Kaplan and Kay, whose work remained an underground classic until it was finally published in (Kaplan and Kay 1994). Eventually, computational linguists interested in describing the wealth of detail present in the phonology and morphology of agglutinative languages got frustrated with the problem of context sensitive parsing, and the practical solution offered by Koskenniemi (1983), propelled both by the Xerox rule compiler (Dalrymple *et al.* 1987) and by Antworth's (1990) PC implementation, became the dominant computational model in the field. To this day, the dominant finite state paradigm is the Xerox regular expression calculus, described in this issue by the **Karttunen** *et al.* paper with syntactic applications in mind. More on the morphological side, but clearly in the same spirit, are the papers by **Tateno** *et al.* on the Japanese lexical transducer and **Koskenniemi**'s paper investigating morphological problems arising in the context of information retrieval.

Finite state syntax, though advocated by a minority throughout the eighties (Ejerhed and Church 1983; Kornai 1985), did not really come in from the cold until the nineties. The present issue offers some prime examples of this work in the papers by **Abney** and **Roche**, who employ finite state methods to describe phenomena, such as light verbs, which were in the tradition of Chomsky (1970) treated as core cases of transformational grammar. The paper by **Vilar** *et al.* describes finite state methods of machine translation, and **Ejerhed**'s paper pushes the envelope even farther, by offering a finite state model of key discourse phenomena. Another important way in which mainstream syntax is impacted by finite state techniques can be called "finite state to the rescue" – the paper by **Schulz and Mikołajewski** describes how constraint-based grammars can be speeded up by finite state methods, and the paper by **Srinivas** shows how corpus-based acquisition of LTAGs is facilitated by finite state techniques.

Perhaps the clearest sign that finite state approaches became part of the mainstream is that they are now subject to the same trends as the rest of computational linguistics. In particular, we see an increased interplay between the statistical and rule-based paradigms in this domain. In some part this is due to the finite state nature of much statistical work (in fact the founding paper of the field, Markov (1913) can be seen in retrospect as a finitary model) but in greater part it is due to an increased awareness on the part of grammar writers that certain aspects of the system, most notably the relationship between the spoken, written, or signed signal and the underlying psychological units, resist characterization in non-statistical terms. An important step in bringing rule-based and statistical work closer is the framework of weighted transducers developed at AT&T/Bell Labs, represented in this issue by **Sproat**'s paper.

Most readers of JNLE are likely to find this special issue a good introduction to current trends in finite state language modeling. But there is an important class of readers that the current selection of papers will leave somewhat dissatisfied: people interested in the mathematical foundations will find only one paper here, by **Bertsch and Nederhof**, which adds to their knowledge of the subject. While it is certainly true that the mathematical theory of (weighted) regular sets and relations is mature, the same can not be said of the algorithmic aspects of the subject, and as the size of the machines grows, the search for more efficient algorithms is likely to intensify. But readers of this special issue should take comfort in the knowledge that finite state methods already offer a degree of efficiency and scalability unmatched by any other technique of natural language engineering.

Acknowledgements

This issue of JNLE could not have come into being without the work of the other workshop organizers, Eva Ejerhed (chair), Frederic Jelinek, and Lauri Karttunen. In addition to them, special thanks are due to the other referees, Salah Ait-Mokhtar, Vagelatos Aristides, Erzsébet Csuhaj-Varjú, Aravind Joshi, Fred Karlsson, Kimmo Koskenniemi, Doug Merritt, Mehryar Mohri, Emmanuel Roche, Richard Sproat, and those who wished to remain anonymous.

References

Antworth, Evan. 1990. *PC-KIMMO: A Two-level Processor for Morphological Analysis.* Texas: Summer Institute of Linguistics.

Chomsky, Noam. 1970. Remarks on nominalization. In Jacobs, R. and Rosenbaum, P. (eds.), *Readings in English transformational grammar* Blaisdell, Waltham, Mass. pp. 184-221

Chomsky, Noam and Halle, Morris. 1968. *The Sound Pattern of English* Harper and Row, New York

Dalrymple, Mary, Kaplan, Ronald M., Karttunen, Lauri, Koskenniemi, Kimmo, Shaio, Sami, and Wescoat, Michael. 1987. Tools for morphological analysis. Center for the Study of Language and Information Report. CSLI-87-108, Stanford

Ejerhed, Eva and Church, Ken. 1983. Finite State Parsing. In Karlsson, F. (ed.), *Papers from the Seventh Scandinavian Conference of Linguistics* pp. 410-432

Jelinek, Frederic. 1997. Language modeling for speech recognition. To appear.

Johnson, Ch. Douglas. 1970 *Formal aspects of phonological representation.* PhD Thesis, UC Berkeley

Kaplan, Ronald M. and Kay, Martin. 1994. Regular Models of Phonological Rule Systems. *Computational Linguistics* **20** (3): 331–378.

Kleene, Stephen C. 1956. Representation of events in nerve nets and finite automata. In Shannon, C. and McCarthy, J. (eds.), *Automata studies* Princeton University Press pp. 3-41

Koskenniemi, Kimmo. 1983. Two-level Morphology. A General Computational Model for Word-Form Recognition and Production. Department of General Linguistics. University of Helsinki, Finland.

Kornai, András. 1985. Natural languages and the Chomsky hierarchy. In: *Proceedings of the 2nd European ACL Conference* pp. 1-7

Markoff, A.A. 1913. Essai d'une recherche statistique sur le texte du roman 'Eugene Onegin'. *Bull. Acad. Imper. Sci. St. Petersbourg* **7**

Mealy, G.H. 1955. A method for synthesyzing sequential circuits. *Bell Systems Technical Journal* **34**: 1045-1079

Miller, George A. and Chomsky, Noam. 1963. Finitary models of language users. In Luce, R. Duncan, Bush, Robert R. and Galanter, Eugene (eds.), *Handbook of mathematical psychology* Wiley, New York. pp. 419-491

Spencer, Henry. 1986. `regexp(3)`. Posted on `mod.sources` **3** (89). See also *A Regular-Expression Matcher.* Ch. 3 in Schumacher, Dale (ed.), *Software Solutions in C* Academic Press

Thompson, Ken. 1968. Regular Expression Search Algorithm. *Communications of the ACM* **11**: 419-422

Wu, Sun and Manber, Udi. 1992. Fast Text Searching Allowing Errors. *Communications of the ACM* **35** (10): 83-91