Comments on Mohri, Pereira and Riley

András Kornai
IBM Almaden Research Center
650 Harry Road, San Jose, CA 95120
kornai@almaden.ibm.com

Throughout the history of computational linguistics the rule-based and the statistics-based approaches appeared as competing rather than complementary threads of research. Even today, many view the success of the Xerox rule-based taggers as a threat to the more statistically oriented taggers, just as a few years ago the IBM statistical approach was viewed as a threat to rule-based systems of machine translation. The historical importance of the AT&T work on weighted transducers lies in the fact that for the first time it makes possible a genuine integration of the two approaches.

Under the heading of "speech-natural language integration" we usually find the shotgun marriage of two completely disjoined systems, each with its own distinct theoretical apparatus and algorithmic building blocks. In contrast, here we find surprisingly smooth integration, both in terms of underlying theory and in terms of shared algorithms. This is a very significant accomplishment, and the main goal of my comments is to situate it as a particular stage of a constant developmental trend towards greater integration. I will ask how much the good sides of the rule-based and the data-based approaches have been preserved, and what, if anything, has been lost by taking this approach. I will also ask how far the present approach can be pushed, and offer some speculative remarks on future directions.

First let me take a clear and unambiguous stance on the rule-based vs. statistics-based debate: rules are better. As a simple illustration, consider Fig. 1 which shows the performance of a bank check OCR system developed by the author [2] under three conditions: using a bigram language model, a finite state grammar, and a combination of the two.

As it is evident from Fig. 1, the rule-based system fares much better than the statistics-based, and in fact the latter adds very little to the performance of a system already containing the former. So the question is not so much an overarching philosophical problem of whether rules are better, but rather the more mundane practical problem of finding the rules. To the extent that the rules, constrains and representations constituting the grammar are devised by grammarians like Quirk and Greenbaum [3], computational linguists can get a free ride, and they should avail themselves of the opportunity. But to the extent that Quirk et al. represent the culmination of an extremely sophisticated descriptive tradition of a singularly deeply researched language, it appears very unlikely that more than a handful languages could be handled in the same fashion.

Again as an illustration, readers are invited to consider

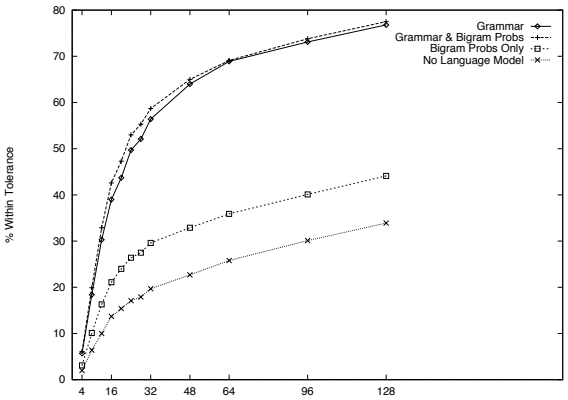

Figure 1. Rules vs. statistics

what kind of grammar they would write for the language of US personal checks before consulting Fig. 2. The regularities of the English numeral system are not hard to capture in a few context free rules, and limiting the dollar amount to four digits will in fact yield a system that can be compiled into a finite state network. But the ideal grammar describing the numerals yields only some of the rules used in the actual grammar, the rest comes from rules dealing with the various types of noise (including the space-filler horizontal line) that we find on checks.

Figure 2. Check grammar

Enoise Lnoise Body Lnoise [**fr**] *Lnoise* **ds** *Lnoise Enoise*
Enoise ⇒ [**ws**] (**bl ws**)* [**bl**]
Lnoise ⇒ ([**ln**] (**ws ln**)* [**ws**]) | ([**hy**] (**ws hy**)* [**ws**])
Body ⇒ *Fourdig* | *Threedig* | *Twodig* | *Onedig*
Onedig ⇒ *Dig* [**ws**] [**dhs**] [**ws**] [**and** | **aps**]
Dig ⇒ **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9**
Twodig ⇒ *Decade*[**ws**] [**hy**] [**ws**] [*Dig*] [**ws**] [**dhs**] [**ws**] [**and** | **aps**]

Now let us take a look at the transducers on the path from acoustic data to sentences. The acoustic observation acceptor O fills only a technical purpose and can be disregarded here. The transducer A from observation sequences

Proceedings of the ECAI 96 Workshop
Extended Finite State Models of Language
Edited by A. Kornai.

to context-dependent phones captures the essence of the relationship between the underlying linguistic unit and the observable signal. In a fuller model, we could in principle decompose it in three parts: A_3 from linguistic units to nerve impulses governing the vocal tract, A_2 from nerve impulses to articulator positions, and A_1 from articulator positions to acoustic observations, with $A = A_1^{-1} \circ A_2^{-1} \circ A_3^{-1}$. Each of these components are governed by rules of biology and physics (only the mapping between cognitive units and nerve impulses has the characteristics associated with linguistic rules, i.e. that they are discrete, fixed, and arbitrary) and all of these rules are expressible as weighted transducers over quantized variables. There is a great deal of work e.g. at Haskins Labs to express these rules in a tabular format, and it would be interesting to see whether the weighted transducer mechanism, which is obviously expressive enough to encode table lookup, could benefit these efforts.

As the authors note, special-purpose context-dependency machinery is commonly found in recognizers, e.g. the triphone mechanism in HTK. The fact that this machinery can be replaced by finite state transducers comes as no surprise, and the approach generalizes well to more complex context-dependent units. However, here we have reached the point where the expertise of the grammarian is minimal, and for the most part we only have the vaguest clues what units to use and what data to pack in them. At the next stage D from sequences of phone labels to a specific word, we have strong empirical evidence that even the most detailed pronunciation dictionaries omit a large number of attested phonetic realizations of most words. Kenyon and Knott [1] are no Quirk and Greenbaum, and it simply does not seem possible to leverage their work the same way. It can be safely predicted that in this domain the statistical approach will reign supreme for many years to come.

Finally there is the language model M, currently presented by the authors as an n-gram model. As the example in Fig. 1 shows, n-gram models are a highly inefficient way of extracting regularities from any domain. So the Xerox program of extracting regularities by grammarians, limited as it may be by the "John Henry argument", remains relevant. What we want is a compact representation of the grammarian's knowledge, fast enough so that alternative formulations can be tested and debugged. Since this knowledge is generally presented conjunctively, intersection remains an essential operation in the creation of fast models. It is not obvious how the Xerox program can be carried to its conclusion using lazy composition methods.

As an example let us consider the old problem of 'readjustment rules' which govern the interaction between syntactic structure and phonological phrasing. Since phrasing triggers a great number of postlexical rules governing e.g. tonal melodies and sentential stress placement, and the latter have wide-ranging autosegmental and even segmental implications, here we have a case where the interaction of every member in the cascade of (weighted) transducers is relevant. For the sake of simplicity consider a sequence of syllables, the leftmost one lexically prespecified as H and on the rightmost one as L. Such a configuration leaves open the possibility of any spreading pattern from HLL...LLL and HHL...LLL to HHH...HLL and HHH...HHL.

Simplifying matters somewhat, this corresponds to the composition $a \circ b$ of two automata a and b. a, corresponing to the segmental tier, is defined by a single loop over tonally arbitrarily specified archisegments. b, corresponing to the tonal tier, is defined by states H and L, with loops over the two states and a unidirectional transition from H to L. Since this last transition can be taken only once, triggered by the appropriate syntactic conditions, we need to intersect the composed automaton with another (possibly very complex) automaton c that encodes the relevant syntactic conditions. If we had lazy addition, taking advantage of a fast multiplication algorithm could be problematic in a process computing $(a+b)*c$. Here the inner operation is composition, the outer is intersection, and in a typical case c would itself be a composition/intersection of automata. In our simplified example, $a \circ b$ can be easily computed offline, but with larger rule systems the issue of intersecting 'lazily given' machines can easily become a significant one.

With the rise of Optimality Theory phonology is increasingly moving toward a style of grammar based on the interaction of very general, typically universal, constraints. To the extent such constraints are non-local it becomes critical for the search space to carry state information in a format very different from that suggested by beam search. It is not that too many alternatives need to be kept open, but rather that these alternatives are not close enough to one another at the interfaces of the uncomposed machines. To put it another way, to leverage the knowledge of the grammarian given in a set of (possibly violable) constraints we need some mechanism for efficient intersection (and subsequent minimization) of machines that are themselves given as cascades, and currently computed only in a lazy manner.

Let me conclude by speculating a bit about the future of the AT&T approach. For the moment, the authors are concentrating on transducers whose range (weight structure) is most appropriate for probabilities or log probabilities. However, valuation semirings of a more discrete character, in particular, valuations in natural numbers corresponding to the degree of constraint ranking and constraint violation, should also be considered. To the extent that our goal is to leverage the information provided by the grammarian, and this information is given to us in terms of ranked violable constraints, we must include operations on weighted automata and transducers not commonly considered, such as restriction to a certain rank, and rank-prioritized intersection.

REFERENCES

[1] John Kenyon and Thomas Knott, *A pronouncing dictionary of American English,* G. & C. Merriam, Springfield MA (1944) McGraw-Hill, New York (1970)

[2] András Kornai, K.M. Mohiuddin and Scott D. Connell, 'Recognition of cursive writing on personal checks', Proc. 5th International Workshop on the Frontiers of Handwriting Recognition, Essex 1996 (to appear).

[3] Randolp Quirk, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik, *A grammar of contemporary English,* Longman, London (1973)