Hybrid Distributions of Strings

James Rogers
Dept. of Computer Science
Earlham College
Richmond, Indiana USA[*]

This talk reports on current work with Jeff Heinz (University of Delaware), bringing together prior results on probabilistic models of language, which promises to lead to hybrid models that are sensitive to both local and long-distance dependencies. As this is work in progress the last third of the talk will discuss preliminary, even speculative, results.

Strictly Local Stringsets

Membership in a Strictly 2-Local (SL_2) stringset (McNaughton and Papert 1971) is determined by the initial symbol, by the pairs of symbols that occur in the string and by the final symbol (independently). Thus they can be specified by giving a set of pairs of symbols drawn from the alphabet augmented by a start and an end marker. The members of the stringset are all and only those in which all substrings of length two (the *2-factors* of the string) are licensed by this set of pairs. Similarly, the stringset can be specified by the set of unlicensed pairs, its set of *forbidden factors*.

These stringsets can be characterized by *Myhill graphs* (Myhill 1957; Lawson 2004), a restricted form of Finite State Automata (FSA) in which the state set is the alphabet plus the endmarkers and there is an edge from one state to another iff they form a licensed pair.

Strictly Local Distributions

Probabilistic Myhill graphs are ordinary probabilistic DFAs (Rabin 1963) based on Myhill graphs. These can model languages under the assumption that the probability of one symbol following another in a string is independent of the portion of the string preceding the pair. This is the Markov property and Strictly 2-Local Distributions are, simply put, the 2-gram models.

Strictly Piecewise Stringsets

The class of Strictly 2-Piecewise (SP_2) stringsets (Rogers et al. 2010) is analogous to the class SL_2 with the exception that membership is determined not by the substrings of the string but, rather, by its *subsequences*, sequences of symbols that occur in the string in order but with arbitrary intervening strings. An SP_2 stringset can be specified by a set of pairs of symbols, in this case not including end markers, in the same way that SL_2 stringsets are specified by sets of licensed or forbidden 2-factors.

The canonical FSA model of an SP_2 stringset is a set of two-state automata each accepting the set of strings in which a forbidden subsequence does not occur. There is a transition from the start state on every symbol of the alphabet, all of which return to the start state except the transition on the initial symbol of the pair, which leads to the second state. The second state has a transition for every symbol except for the second symbol of the pair, all of which return to the second state. Both states are accepting. The string is rejected iff the second symbol of the pair is encountered while in the second state.

The SP_2 stringset is the intersection of the stringsets accepted by these factored automata.

Strictly Piecewise Distributions

These sets of factored automata serve as the structural framework of probabilistic models of language under the assumption that the probability of one symbol occurring somewhere in the string following another is independent of the probability of that symbol occurring following any other symbol (Heinz and Rogers 2010). We call this the Piecewise property. Under this assumption the probabilities, for a given stringset, of the transitions in the product automaton can be computed from the probabilities of the transitions in the factor automata.

SL+SP Stingsets

The SL+SP stringsets are the intersections of SL and SP stringsets (Heinz and Rogers 2013). Hence they capture the co-occurrence of (strict) local and (strict) long-distance constraints. This class has proven to be particularly relevant to phonotactics: 82% of the patterns in Heinz's catalog of metrical stress in known languages can be characterized as the co-occurrence of constraints of this sort. All of the patterns in Heinz's catalog, with the exception of four patterns, all of which involve an alternation that does not show up in the surface form, can be characterized by the co-occurrence of (non-strict) Locally Testable (LT) and SP constraints (Wibel, Rogers, and Heinz).

[*] Joint work with Jeff Heinz, University of Delaware

Hybrid Distributions

This hybrid class, then, may provide the structural framework for probabilistic models which are based on both local and long-distance phenomena. The underlying assumption of such distributions combines the Markov property with the Piecewise property along with the assumption that the distributions of 2-factors and 2-pieces are independent. This is quite a strong assumption, requiring, for example, that the probability of a symbol eventually following another is independent of the probability that it immediately follows it. It remains to be seen whether probabilistic models based on this assumption can be faithful enough to be useful.

References

Heinz, J., and Rogers, J. 2010. Estimating strictly piecewise distributions. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 886–896. Uppsala, Sweden: Association for Computational Linguistics.

Heinz, J., and Rogers, J. 2013. Learning sub-regular classes of languages with factored deterministic models. In Kornai, A., and Kuhlmann, M., eds., *Mathematics of Language*. Association for Mathematics of Language.

Lawson, M. V. 2004. *Finite Automata*. Chapman and Hall/CRC.

McNaughton, R., and Papert, S. 1971. *Counter-Free Automata*. MIT Press.

Myhill, J. 1957. Finite automata and the representation of events. Technical Report 57-624, Wright Air Development Command.

Rabin, M. O. 1963. Probabilistic automata. *Information and Control* 6:230–245.

Rogers, J.; Heinz, J.; Bailey, G.; Edlefsen, M.; Visscher, M.; Wellcome, D.; and Wibel, S. 2010. On languages piecewise testable in the strict sense. In Ebert, C.; Jäger, G.; and Michaelis, J., eds., *The Mathematics of Language: Revised Selected Papers from the 10th and 11th Biennial Conference on the Mathematics of Language*, volume 6149 of *LNCS/LNAI*. FoLLI/Springer. 255–265.

Wibel, S.; Rogers, J.; and Heinz, J. Factoring of stress patterns. In preparation.