Comments on Jelinek

Mehryar Mohri
AT&T Research
600 Mountain Avenue
Murray Hill, 07974 NJ
mohri@research.att.com

Speech understanding systems tend to use grammars that are similar to those utilized for written language. They are often used both to put restrictions on the appearance of some sentences and to parse acceptable sentences. But speech understanding systems depend on recognition. In speech recognition, grammars are not used to parse sentences, nor to put restrictions on the appearance of sentences. Speech is a noisy data, and a priori any sentence can be associated with a given utterance. The probability of such an association is of course different for each pair of utterance and sentence. That is why speech recognition systems use *grammars* based on statistical methods, in the case of words, language models, that can be used to assign different probabilities to each sentence[1] They constitute one of the essential components of most systems other than those with a limited vocabulary.

Frederick Jelinek gives here a clear and compact overview of three different approaches used to construct language models in speech recognition:

- Models based on n-grams,
- Models using decision trees, and,
- Models based on the maximum entropy.

It is a remarkable fact that all these approaches lead to a finite-state language model, a *weighted automaton* or a finite-state transducer from strings to weights [1]. One could think of more complex statistical models such as probabilistic context-free models. But, such models are not typically used in practice because they need to be lexicalized to contain adequate information, the corresponding methods are still experimental. The computational efficiency of those models is also a concern when building real-time systems. Finite-state machines lead to much more efficient programs.

It is also worthwhile to point out that in spite of the number of years and the number of works dedicated to the construction of language models, the first approach is still the most popular and the one widely used in speech recognition systems. Many refinements have been introduced for the construction of n-grams. Some have been successfully used to build pentagrams. But most *subtleties* are related to smoothing techniques. Jelinek mentions the main ones. Once again, in most speech recognition systems the Katz's method is used. That uniformity of the approaches adopted in practice might suggest that researchers do not consider the language model to be the main problem in speech. This is not the case though.

From that respect, it is important to investigate new approaches and Jelinek describes two interesting ones. The approach based on decision trees is often used in building acoustic models. The main problem in using that approach is, as pointed out in the paper, the choice of the parameters or the questions. What are the good questions from the linguistic and computational point of view? Since the size of the tree is exponential in the number of alternatives for a given question, one needs to limit the number of alternatives. An interesting recent work makes the use of that approach more attractive by introducing a method to construct a weighted finite-state transducer directly from decision trees [4], by first transforming the tree into a set of context-dependent rules and then using a weighted rewrite rule compiler to construct an equivalent weighted transducer [3].

The last approach described is the maximal entropy one: it allows one to construct a trigram language model in which probabilities verify a set of constraints usually represented by characteristic functions and such that the probability defined have maximal entropy. The method depends on thresholds defining the characteristic functions. The model has some flexibility since the thresholds can be used to adjust the resulting size of the model. The main problem is to compute a set of coefficients $lambda_i$ such that the constraints be verified. The paper describes an efficient solution to that problem which is linear in the size of the vocabulary.

One can hope that the new methods described by Jelinek will be used in practice, or that perhaps they encourage researchers to investigate constructions more refined than the basic ones for building n-grams. More refined linguistic methods could also be investigated. Lexicon-grammars such as those of LADL with a large lexical coverage of language, in particular the local grammars [2], could perhaps be used to increase the precision of language models, if they could be combined with statistical methods.

[1] This does not mean of course that language models cannot be used in non-speech applications.

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

REFERENCES

[1] Jean Berstel, *Transductions and Context-Free Languages*, number 38 in Leitfäden der angewandten Mathematik and Mechanik LAMM, Teubner Studienbücher, Stuttgart, Germany, 1979.

[2] Maurice Gross. The use of finite automata in the lexical representation of natural language. *Lecture Notes in Computer Science*, 377, 1989.

[3] Mehryar Mohri and Richard Sproat, 'An efficient compiler for weighted rewrite rules', in *34th Annual Meeting of the Association for Computational Linguistics*, San Francisco, California, (1996). University of California, Santa Cruz, California, Morgan Kaufmann.

[4] Richard Sproat and Michael Riley, 'Compilation of Weighted Finite-State Transducers from Decision Trees.', in *34th Annual Meeting of the Association for Computational Linguistics*, San Francisco, California, (1996). University of California, Santa Cruz, California, Morgan Kaufmann.