NEW APPROACHES TO STOCHASTIC MODELING OF SPEECH

Joseph Picone

Institute for Signal and Information Processing
Mississippi State University
Mississippi State, Mississippi 39762
picone@isip.msstate.edu

Mari Ostendorf

Department of Electrical, Computer, and Systems Engineering
Boston University
Boston, MA 02215
mo@raven.bu.edu

ABSTRACT

Hidden Markov Models and n-gram language modeling have been the dominant approach in continuous speech recognition for almost 15 years. Though successes have been well-documented, fundamental limitations of this paradigm surface at both the acoustic and language modeling ends of the speech recognition problem. Although acoustic models based on linear statistical assumptions have led to steadily improved performance on speech collected in benign environments, they are still sorely lacking on spontaneous data encountered in the field. Similarly, robust parsing of dialogs and unconstrained man-machine communications is a serious problem for today's technology.

In this session, we attempt to stimulate a discussion on new approaches in statistical modeling. Researchers from both inside and outside of the speech community are invited to present new perspectives on how complex behavior can be modeled in a parsimonious manner. Our panel discussion will attempt to identify and debate a handful of promising new directions in statistical modeling of speech.

1. INTRODUCTION

Currently, the most successful speech recognition systems use detailed models of local context with large numbers of parameters trained in a limited domain. In acoustic modeling, context-dependent hidden Markov models (HMMs) [1,2] have become a standard approach for handling the variability due to local phonetic context, where local context may be a window of 3-5 phones. In language modeling, trigrams [3] have dominated the field, with major improvements coming from use of higher-order n-grams. For problems where training data has been steadily increasing these models show steady improvement, as demonstrated in the NAB benchmarks where word accuracy rates of less than 10% have been achieved on an open vocabulary task [2]. However, the same technology has provided minimal advances on less constrained tasks, specifically on the spontaneous conversational speech found in the Switchboard corpus where error rates have not moved much beyond 50% word accuracy [4] in the past two to three years. In addition, commercial researchers often observe error rate increases of factors of 3-4 when moving lab technology to the field.

Of course, both trigram language models and hidden Markov acoustic models have made tremendous contributions to progress in speech recognition. They clearly established the value of automatic training algorithms and the usefulness of Markov assumptions in simplifying both recognition and training complexity. In addition, the successes provide important evidence that local context — neighboring phonemes in acoustic modeling and neighboring words in language modeling — provides the most important information for speech recognition. Conditioning on local context seems to be an important attribute of a good statistical model. The question raised here is whether sufficient progress can be made simply by increasing the number of parameters in these models. Adaptation certainly helps improve performance, but it is predicated on reasonably accurate baseline performance. Progress in speech recognition is ultimately limited by the sophistication of statistical models, and current technology is unlikely to provide the capability for computers to really converse with humans. New models are needed to better capture variability at a local level and/or to model trends operating at a higher level.

Nonlinear systems theory has been an active area of research in the last twenty years [6] in the field of dynamics. What is new is that it promises to be an active area of engineering in the next twenty years as a new wave of mathematics begins moving from the laboratory to the field. Much as linear system theory provided tools for scientists to analyze classes of problems previously thought too complicated, nonlinear system theory offers hope of providing tools to unlock the mysteries of a wide range of important biological signals such as speech.

Similarly, computational research in language has spanned decades. In the late 1950s, a hierarchy of grammatical formalisms [5] was defined in an attempt to document the complexity of language. As HMMs were introduced in speech recognition, great excitement was generated by the fact that both acoustic models and language models could be represented as state machines. Researchers were quick to see, however, that this was simply the first step in representing the entire speech recognition problem as a formal language theory problem. HMMs were shown to be equivalent to regular grammars, and shown to simply be one step in a progression towards context-sensitive grammars. Today, we find systems routinely implementing context-free grammars and left regular grammars. True context-sensitive grammars, however, have so far been impractical for speech recognition and understanding applications. In addition, experiments in understanding spontaneous speech (e.g. in the ATIS task [2]), have shown that conventional parsing techniques are ill-suited to processing spontaneous spoken language and various robust parsing algorithms are now being explored.

In sum, research in both acoustic and language modeling currently benefits from the power of context-sensitive statistics, but both are also limited in not moving beyond the local level. Too many parameters are dedicated to the local structure at the expense of capturing global structure. As noted in [6], “Knowing the microscopic laws of how things move still leaves us in the dark as to their larger consequences.” One of the attractions of nonlinear systems is the hope of modeling the coarse behavior of a system in which a detailed analysis is not required, a very common problem in statistical mechanics. Similarly, one of the attractions of grammatical language models is the potential for capturing the higher level structure inherent to language. It is clear that our current formalisms are not adequate for the difficult recognition tasks at hand. Here, we take a look at some new directions that may offer a means of overcoming limitations of existing statistical models.

2. SESSION OVERVIEW

This session consists of four invited panelists:

- Simon Haykin, McMasters University
  “Chaotic Statistical Modeling”
- Tony Robinson, Cambridge University
  “Some New Uses of EM in Acoustic Modeling”
- Ted Briscoe, Cambridge University
  “Language Modeling or Statistical Parsing?”
- Fred Jelinek, Johns Hopkins University
  “A Context-Free Headword Language Model”

The panelists were selected to provide perspectives on two key dimensions of the speech understanding problem: acoustic modeling and language modeling. On each of these topics, we have included a speaker drawn from outside the normal speech research community, and a speaker representing a somewhat more mainstream viewpoint.

The first two talks deal with the issue of nonlinear acoustic modeling. The piecewise constant model has been a staple of digital speech processing since the early 1970’s. A multivariate Gaussian model of observation vectors has been employed in hidden Markov model-based speech recognition systems since the early 1980’s. Though neural network-based approaches have been researched since the mid-1980’s, only recently has the performance of such models rivaled conventional technology.

Simon Haykin suggests a new approach to signal modeling based on chaotic signals. His research is representative of a new body of science devoted to the application of nonlinear dynamics to conventional classification problems. Classification of signals into deterministic and stochastic ignores an important class of signals, known as chaotic signals, that are deterministic by nature yet random in appearance. While direct modeling of the speech signal as output from a nonlinear system has not proven to provide enhancements over conventional analyses, recent research suggests these techniques are applicable to the statistical modeling problem that is the core of the acoustic modeling problem. In his talk, Haykin advocates an architecture that employs neural networks to perform the actual prediction/ detection task. This is not unlike many hybrid speech recognition systems that now use a combination of hidden Markov models and neural networks.

In a companion talk, Tony Robinson discusses issues in acoustic modeling in the context of connectionist/HMM systems, which use neural networks to estimate posterior distributions for HMM states, and can be thought of as a non-linear extension of current HMM technology. While his general approach is based on non-linear models, the theme of his talk is new applications of one of the most powerful tools behind HMM technology: the expectation-maximization (EM) algorithm [7]. Robinson examines the notion of a hidden component of the process, e.g. the state in an HMM. He explores how this state, which is currently used to capture contextual and temporal phonetic variability, can improve aspects of speech recognition from feature extraction to posterior distribution modeling to channel or outlier (goat) identification.

The subsequent two talks deal with issues related to the language modeling problem, which many people believe will be the source of most of the improvement in speech understanding systems in the next few years. We have seen many attempts at improving recognition through more sophisticated language modeling, from higher-order and variable-order n-grams to the introduction of grammatical structure, but such techniques have generally resulted in modest improvements in performance at the expense of significant increases in complexity. However, new work aimed at combining the advantages of grammatical structure and local context are emerging in various forms of lexicalized grammars. This combined approach may offer the most hope for performance gains, and one theme of the two language modeling talks is lexicalization and its practical implementation.

Ted Briscoe suggests that interpretation of text requires a framework beyond n-gram models, due to a need of such systems to evaluate the relative likelihoods of complex grammatical relationships, which are hierarchical and therefore not easily captured by n-gram models. He advocates the use of statistical parse selection models over stochastic context free grammars, but quickly notes that integration of estimate-maximize (EM) techniques into these formalisms is challenging. Such formalisms do not lend themselves to the treatment of conditional probabilities and maximum likelihood calculations involving ambiguous-in-time candidate partial parses. Research to-date in formalisms beyond CFGs has been disappointing. Yet, it is clear that such formalisms are needed to deal with the complex language models required for spontaneous dialogues.

In a companion talk, Fred Jelinek introduces lexicalized stochastic context-free language model that takes advantage of a parser to define the phrase structure of a word string. The authors use the notion of a phrase headword to relate non-terminals directly to lexical items, and use the given parse structure to reduce the cost of computing the probability of a word string. The problem of sparse data in parameter estimation is addressed by defining word classes as would be used in a class grammar, but here the classes simply define a smoothing hierarchy. This approach is representative of a growing trend towards lexicalized grammars.

3. DISCUSSION

Perhaps the most important aspect of this session will be the panel discussion held after the plenary talks. Some of the issues we feel are an outgrowth of the papers presented in this session include:

- What will be the impact on the number of parameters required in a speech recognition system based on nonlinear statistical models?

  If more parameters are required, then the benefits of such an approach might vanish for small training sets. If the number of parameters decreases, perhaps sensitivity to speaker or channel will increase.

- If linear models work well for variations within a phone as long as the time span is sufficiently small, might that argue for the use of non-linear models to represent the hidden component?

  New and improved spectral estimation techniques have often not proven to provide substantial gains in recognition performance, perhaps leading us to believe the hidden component of the acoustic model is the area needing a better statistical model. However, non-linear techniques are also motivated by articulatory models, where non-linearities are usually placed at the output level rather than embedded within the internal structure of the model.

- Are the language models presented here only practical in an n-best rescoring framework, or can we envision using them to reduce the search space? Are there better ways to integrate new language models with acoustic modeling to produce more efficient recognition systems?

  When recognition performance is poor, or the search space is large, n-best outputs can be quite limiting, forcing one to sift through large numbers of competing hypotheses that look very similar. New language models must find a way to limit the number of hypotheses passed to higher levels of the system.

REFERENCES

1. S. Young, “Large Vocabulary Continuous Speech Recognition: A Review,” presented at the 1995 IEEE Automatic Speech Recognition Workshop, Snowbird, Utah, U.S.A., Dec. 1995.
2. D.S. Pallett, *et. al.*, “1994 Benchmark Tests for the ARPA Spoken Language Program,” in *Proceedings of the ARPA Spoken Language Systems Technology Workshop*, Austin, Texas, USA, January 1995, pp. 5-36.
3. F. Jelinek, R. Mercer and S. Roukos, “Principles of Lexical Language Modeling for Speech Recognition,” in *Readings in Speech Recognition*, ed. A. Waibel and K.-F. Lee, Morgan Kaufmann Publishers, 1990.
4. A summary of recent SWITCHBOARD results are available at the URL: http://cspjhu.ece.jhu.edu.
5. J.R. Deller, J.G. Proakis, and J.H.L. Hansen, *Discrete Time Processing of Speech Signals*, MacMillan, New York, New York, USA, 1993.
6. H.O. Peitgen, H. Jurgens, and D. Saupe, *Chaos and Fractals: New Frontiers of Science*, Springer-Verlag, New York, New York, USA, 1992.
7. A.P. Dempster, N.M. Laird and D.B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society*, Vol. 37, No. 1, pp. 1-38, 1977.

CHAOTIC SIGNAL PROCESSING

Simon Haykin

Communications Research Laboratory
McMaster University
Hamilton, Ontario, Canada
haykin@synapse.crl.McMaster.CA

SUMMARY

Traditionally, signals have been classified into two basic types: deterministic, and stochastic. This classification ignores an important family of signals known as chaotic signals, which are deterministic by nature and yet exhibit many of the characteristics that are normally associated with stochastic signals.

In this talk, we begin by reviewing some important aspects of nonlinear dynamics. This would then naturally lead into a discussion of chaotic systems, how they arise, physical phenomena that are known to be chaotic, and their practical applications.

The second half of the talk will be devoted to the characterization of chaotic signals and the theory of embodology, with emphasis on time series analysis. Specifically, we will describe the following notions:

- Takens' embedding theorem
- Attractor dimension, and the correlation dimension
- Minimum embedding dimension, and its estimation using the method of false nearest neighbors
- Lyapunov spectrum, and its estimation
- Recursive prediction, and how to implement it using neural networks

SOME NEW USES OF EM IN ACOUSTIC MODELING

Tony Robinson

Engineering Department
University of Cambridge
Cambridge, England
ajr@softsound.com

SUMMARY

This talk will raise some problems with the current techniques used in acoustic modeling and suggest some directions for future research. Firstly the connectionist/HMM system known as ABBOT will be briefly introduced. The talk will then progress to suggest a series of new and largely untested applications of the EM algorithm in acoustic vector and acoustic model estimation for automatic speech recognition. These topics are under investigation at Cambridge University and it is hoped that they will contribute to the ABBOT system in the future.

It is acknowledged that the current acoustic vectors used in speech recognition systems are a poor representation of the speech signal. This is clear from speech coding work whereby a standard LPC coder (e.g. LPC10e) may produce unintelligible output in the case of certain speaking styles or mild background noise.

Drawing from speech coding, we can aim to model the parameters of source-filter model such as LPC. In such a model the source is Gaussian white noise or an impulse train. In conventional applications the LPC filter parameters of voiced speech are estimated assuming the white noise source, but it has been shown that an application of the EM algorithm can provide a maximum likelihood estimate to both the LPC parameters and the excitation parameters (the period and phase of the impulse train) [1].

Drawing from speech perception we know that formant locations are a important to vowel identity and that formant frequencies are determined by vocal tract length and are speaker dependent. The simplest speaker invariant parameter is a formant ratio. However, we are not close to incorporating this knowledge in current ASR system as they generally work in the power spectra or cepstral domain. As start in this direction is to estimate the power spectral density as a Gaussian mixture and then to model the trajectories of the Gaussian means [2].

Currently HMMs model acoustic vector densities. We have shown that statistical models of posterior probabilities can also be used [3,4]. However, the conversion of the posterior probabilities to likelihoods involves some approximations which means that, as currently implemented, the training algorithm is not an EM algorithm. These approximations aside, we have shown that we can train on posterior probabilities and that this results in better models over the Viterbi training [5].

It is interesting to consider the connectionist architecture within the EM framework. We consider each unit as estimating an indicator variable which has values of “fire” of “not fire”. We can estimate the MAP probability of firing if we know both the input and the output to the network [6]. Although this work is currently computationally constrained by an exhaustive search it does propose approximations applicable to large networks or the use of Monte-Carlo sampling.

A recent improvement has been the modeling of context dependent phones [8]. Here we assume an indicator variable not only for the phone class at a given time but the phone context given the phone class. We have been able to use connectionist models to estimate this variable which has resulted in improved speech recognition accuracy and speed of decoding.

Another promising candidate for acoustic modeling is the hierarchical mixture of experts [7]. This is essentially a decision tree with a probability of branching associated with each node. The EM algorithm may be used to reestimate the parameters of the system. There are several practical aspects of this architecture that need to be addressed before it can be applied to large speech tasks [9]. The HME can either be applied as a static pattern classifier and a Markov model used to model the dynamics in much the same way as connectionist/HMM hybrid systems or the dynamics can be directly incorporated.

Finally, an acknowledged problem in speech recognition is that some speakers are much easier to recognize than others. Out of ten speakers in a unlimited vocabulary read speech evaluation it is not uncommon for the best speaker to have an order of magnitude lower error than the worst speaker. Hence the overall error rate is dominated by a few outliers (the goats). A proposed “EM” solution to this problem is to label every speaker with an indicator variable (sheep or goat) and use the observed recognition rate to estimate the probability of being a goat. By weighting the training set by these probabilities it is expected that the available modeling power can be better used and the expected error rate decreased.

Another viewpoint on the same scenario is that the goatiness factor is determined by the channel conditions. We consider broadcast speech as a major source of acoustic data for future speech systems. By considering the confidence that the decoded speech came from a clean source rather than a dirty source we hope to filter the unending supply of broadcast speech and train in proportion to the sequential MAP estimate of sheepishness. We hope that this will liberate us from the very significant resources required to construct today’s speech corpora and hence result in significantly better speech systems.

REFERENCES

1. Burshtein 1990: “Joint maximum likelihood estimation of Pitch and AR parameters using the EM algorithm”, pages 797-800, *Proceedings of ICASSP 90*, IEEE.
2. Zolfaghari and Robinson 1995: Cambridge University Engineering Department Technical Report.
3. Bourlard and Morgan 1994: *Continuous Speech Recognition: A Hybrid Approach*, Kluwer Academic Publishers.
4. Robinson Hochberg and Renals 1995: “The use of recurrent networks in continuous speech recognition”, chapter 19 in *Automatic Speech and Speaker Recognition - Advanced Topics*, Editors C. H. Lee, K. K. Paliwal and F. K. Soong, Kluwer Academic Publishers.
5. Senior and Robinson 1996: “Forward-backward retraining of recurrent neural networks”, *Advances in Neural Information Processing Systems*, Vol. 8, Morgan Kaufmann.
6. Cook and Robinson 1995: “Training MLPs via the Estimation-Maximisation algorithm”, *Proceedings of the IEE conference on Artificial Neural Networks*.
7. Jordan and Jacobs 1994: “Hierarchical mixtures of experts and the EM algorithm”, *Neural Computation*, Vol. 6, pp. 181-214.
8. Kershaw, Hochberg and Robinson 1995: “Context-Dependent Modeling in the ABBOT LVSCR System,” presented at the 1995 IEEE Automatic Speech Recognition Workshop, Snowbird, Utah, U.S.A., Dec. 1995.
9. Waterhouse and Robinson 1995: “Pruning and Growing Hierarchical Mixtures of Experts”, *Proceedings of the IEE conference on Artificial Neural Networks*.

LANGUAGE MODELING OR STATISTICAL PARSING?

Ted Briscoe

Computer Laboratory
University of Cambridge
Cambridge, England
ejb@cl.cam.uk

SUMMARY

In *language modeling*, a corpus of sentences is treated as a set of observed outputs of an unknown stochastic generation model, and the task is to find the model which maximises the probability of the observations. When the model is non-deterministic and contains hidden states, the probability of a sentence is the average probability of each distinct derivation (sequence of hidden states) which could have generated it. For speech recognition, language modeling is an appropriate tool for evaluating the plausibility of different possible continuations (word candidates) for a partially recognized utterance. For speech or text understanding the derivations themselves are crucial to distinguish different interpretations. For example:

(a) Charlotte is playing with her rabbit

(b) That rabbit gets played with a lot

(c) ?Charlotte's father gets played with a lot

(d) (S(NP Charlotte) (VP is (VP (V playing with)
(NP her rabbit))))

(e) (S(NP Charlotte) (VP(VP is (VP playing))
(PP with (NP her rabbit))))

In (a), there is an ambiguity between an interpretation in which Charlotte is playing accompanied by her (pet) rabbit and one in which her plaything is a (toy) rabbit. In the latter case *play with* is best analyzed as a phrasal verb and *the rabbit* as a direct object, predicting for example the possibility of passivization in (b).

However, given the former interpretation, *with* is best analyzed as introducing an adverbial prepositional phrase modifier of the verb *play* predicting the accompaniment interpretation of *with* (as one possibility). The oddity of (c), in which we are forced to interpret Charlotte's father as a plaything, is then a consequence of the fact that passivization only applies to direct object noun phrases and not prepositional phrases.

From the perspective of speech recognition, the alternative derivations for (a) are not as important as the relative likelihood with which *rabbit* may follow *play with* (though accurate assessment of the likelihood with which a noun denoting a plaything rather than playmate will be followed by *is/was/gets/...played with* might well require both modeling the distinct derivations and passivization.)

For interpretation, the relative likelihood of the distinct derivations is crucial. Furthermore, the derivation must encode hierarchical consistuency (i.e. bracketing) to be useful. Thus, for (a) we need to know whether the preposition *with* combines with *play* to form a phrasal transitive verb (d) or whether it combines with h*er rabbit* to form a prepositional phrase which in turn combines with the intransitive verb *play* (e), because recognition of which verb we are dealing with determines the difference of interpretation.

N-gram models cannot directly encode such differences of hierarchical organization which is why most stochastic approaches to text understanding have employed stochastic context-free grammars (SCFGs) or feature/unification-based abbreviations of them. Within this framework it is possible to treat the grammar as a language model and return the most likely derivation as the basis for interpretation.

However, there are a number of problems with this approach. Firstly, SCFGs and their feature-based variants associate a global probability with each grammar rule rather than a set of conditional probabilities that a given rule will apply in different parse contexts. A number of experiments by different groups have independently confirmed that modeling aspects of the parse context improves the accuracy with which the correct derivation is selected by up to 30%. Most researchers are now experimenting with statistical parse selection models rather than language models within the text understanding community.

Secondly, SCFGs and close variants, unlike n-gram models, are not easily lexicalized, in the sense that the contribution of individual words to the likelihood of a given derivation is

not captured. Those models which are lexicalized use word forms rather than word senses to condition the probability of different derivations. It is likely that considerable improvement could be obtained by utilizing broad semantic class information (such as that *rabbit* can denote an animal or toy) to evaluate the plausibility of different predicate-argument combinations.

Selecting a correct derivation is only one aspect of the problem of robust practical parsing for text or speech understanding. Another bigger problem is ensuring that the grammar covers the (sub)language. Language models offer one potential solution to this problem, as estimation-maximisation (EM) techniques can be utilized not only to find the (locally) optimal probabilities for grammar rules but also to find the optimal set of rules (by, for example, removing rules re-estimated to some “floor” threshold probability). However, it is not clear that EM techniques can be coherently utilized in this way with statistical parse selection models which cannot be interpreted as language models.

In the talk I will discuss the issues of parse selection vs. language models, lexicalization of models and integrated approaches to statistical rule induction as well as ranking in further detail.

A CONTEXT FREE HEADWORD LANGUAGE MODEL

Ciprian Chelba, Anna Corazza, and Frederick Jelinek

Center for Language and Speech Processing
Johns Hopkins University
Baltimore, Maryland, USA
{chelba, corazza, jelinek}@cspjhu.ece.jhu.edu

SUMMARY

This is an attempt to base a language model on a context-free lexicalized grammar. A language model must be statistical, and therefore simple enough so its parameters can be estimated from data. Since it should reflect message content, it must be lexical.

The lexical non-terminals will be related directly to the words belonging to a vocabulary V and will correspond to the intuitive notion of *headwords*. Headwords are thought to have *inheritance* properties, and so the production rules should basically have the form (s denotes the unique sentence non-terminal located at the root of the tree)

$$\begin{aligned} s &\to H \\ H &\to HG \\ H &\to GH \\ H &\to v(H) \end{aligned} \tag{1}$$

In Eq. 1, $v(H)$ denotes the unique word v which has the *same* name as the headword H.

Figure 1 illustrates a possible derivation of the sentence *Her step-mother, kissing her again, seemed charmed* [1]. Note that attached to the root node of the tree is the headword corresponding to the main verb *seemed* of the sentence.

It is interesting to observe that a *bigram* language model can be considered to have the special headword form

$$\begin{aligned} s &\to H \\ H &\to v(H)G \\ H &\to v(H) \end{aligned} \tag{2}$$

Comparing Eq. 1 with Eq. 2, we see that the general headword language model defined by Eq. 1 has essentially the same parameter complexity as the *bigram* language model. A headword language model whose power could be compared to that of the *trigram* language model would have the structure

$$\begin{aligned} s &\to H \\ H\angle LB(H) = F &\to HG \\ H\angle RB(H) = F &\to HG \\ H\angle LB(H) = F &\to GH \\ H\angle RB(H) = F &\to GH \\ H\angle LB(H) = F &\to v(H) \\ H\angle RB(H) = F &\to v(H) \end{aligned} \tag{3}$$

where the notation $H\angle LB(H) = F \to HG$ means that F is the *left brother* of H and that HG is generated from H. Naturally, $RB(H) = F$ means that F is the *right brother* of H.

We use the automatic transformational (AT) parser of Brill [2] to parse a large amount of text and thus provide a basis, in conjunction with *headword inheritance rules*[1], for the collection of headword production statistics. The availability of an AT parser solves in principle the problems of scarcity of data and of transportability. Any domain with sufficient text data provides sufficient parse data.

A language model is a device that provides to the recognizer the values of the probabilities $P(w_k | w_1, \ldots, w_{k-1})$ predicting the k^{th} word given the hypothesis about the string of the preceding $k-1$ words.

[1] That is rules of the type: *the headword of a simple nounphrase is the last noun;* or *the headword of a verb phrase is its first verb*, etc. Such rules have been derived by linguists [3] for use in the IBM statistical direct parser [4]. Headword inheritance rules can also be derived automatically from data by use of information theoretic principles.

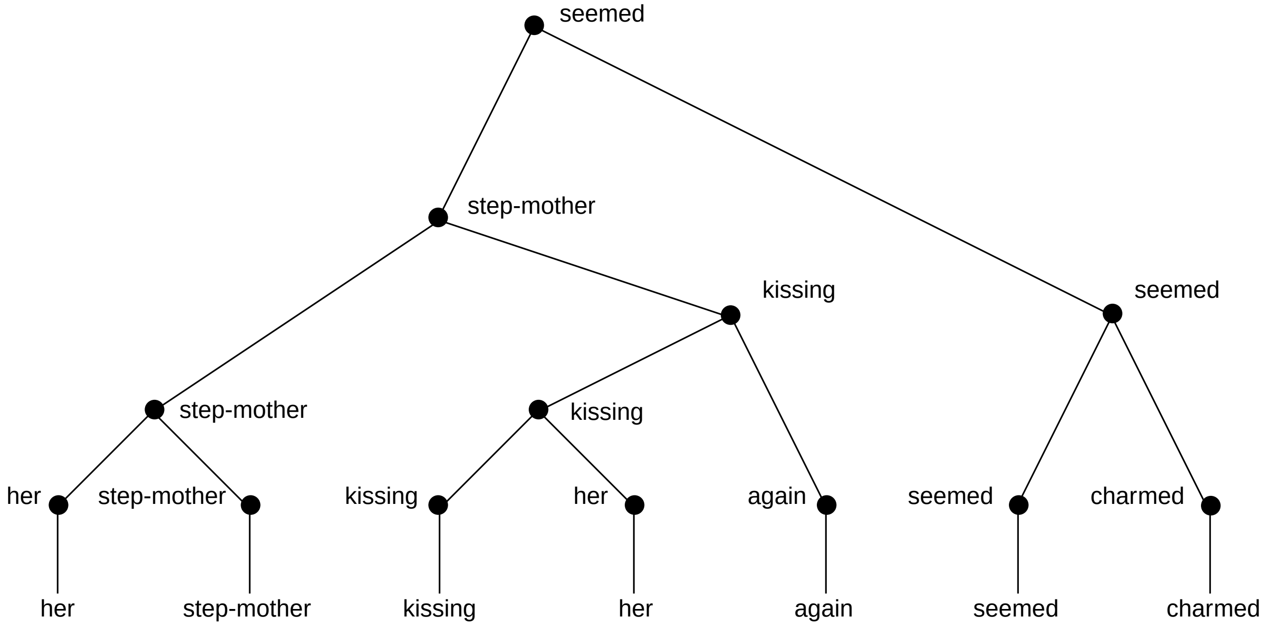

Figure 1. An illustration of a possible derivation of the sentence “Her step-mother, kissing her again, seemed charmed.”

For context free grammars these probabilities can be computed by the method of Jelinek and Lafferty [5]. The headword language model is particularly suited to N best *resolution* when we use the hypothesis that the word strings were produced by the process corresponding to the parse specified by the AT parser applied to these strings. The computation of such a probability involves *one* derivation only and is therefore linear with the length of the sentence.

Clearly, a large amount of parsed data is necessary to adequately estimate headword production probabilities. In fact, many more productions of the types $H \to HG$ and $H \to GH$ will have non-zero probabilities than would bigrams $P(v(G)|v(H))$ of successive words $w_{i-1} = v(H), w_i = v(G)$. The reason is that in $H \to HG$ and $H \to GH$ the headwords H and G may correspond to words that are quite separated in the text from each other.

Therefore, we must smooth efficiently the relative frequencies obtained in the collection process. Let $c(H)$ denote an appropriate *class* of the headword denote an appropriate class of the headword H. Then the appropriate formula is

$$\begin{aligned} P(H \to HG) \;=\; & \lambda_1 f(H \to HG) + \\ & \lambda_2 f(H \to Hc(G)) + \\ & \lambda_3 f(c(H) \to c(H)c(G)) \end{aligned} \qquad (4)$$

We have derived the classes $c(H)$ based on the method of Mercer [6] applied to parse trees.

REFERENCES

1. Henry James: *What Masie knew*, Penguin Books, New York, 1980
2. Eric Brill: *A corpus-Based Approach to Language Learning*, Ph.D. dissertation, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, 1993.
3. E. Black, personal communication.
4. F. Jelinek, et. al.: Decision tree parsing using a hidden derivation model, in *Proceedings of the ARPA Spoken Language Systems Technology Workshop*, Austin, Texas, USA, January 1995.
5. F. Jelinek and J. Lafferty: Computation of the probability of initial substring generation by stochastic context free grammars, *Computational Linguistics*, Vol. 17, pp. 315 - 324, 1991.
6. P.F. Brown, et. al.: Class-based n-gram models of natural language, *Computational Linguistics*, Vol. 18, No. 4, pp. 467 - 480, December 1992.