SOME RESULTS ON STOCHASTIC LANGUAGE MODELLING

Renato De Mori and Roland Kuhn

School of Computer Science
McGill University
3480 University St.
Montreal, Quebec, Canada H3A 2A7

ABSTRACT

The paper will discuss three issues. The first is the derivation of precise probability scores for partial hypotheses containing islands, in the context of a Stochastic-Context-Free-Grammar (SCFG) for Language Modeling (LM). The second issue is the possibility of adding a cache component to a LM. This component alters the expected probability of words to reflect the speaker's patterns of word use. Finally, the idiosyncratic properties of dialogue are being studied; this work will indicate how knowledge about the discourse state can be incorporated into the LM and into the semantic component.

ISLAND-DRIVEN PARSING

Language Modeling and Theories with Islands

Automatic Speech Understanding (ASU) is based on a search process that generates partial interpretations of a spoken sentence called *theories*; theories are scored on the basis of a likelihood $L = O(\Pr(A \mid th)\Pr(th))$. We are interested in the computation of $\Pr(th)$ when th is a partial interpretation of a spoken sentence generated by a Stochastic Context-Free Grammar (SCFG) G_s. A recent report [2] reviews this problem and gives interesting results.

The most popular parsers used in Automatic Speech Recognition (ASR) generate new theories in a left-to-right fashion. To score the theories generated by these parsers, the probability of all parse trees generating the first p words of a sentence must be computed; the appropriate algorithms are given in [8].

Parsers that are "island-driven" proceed outward in both directions from *islands* of words that have been hypothesized with high acoustic evidence. Interesting island-driven parsers have been proposed by [13], [12], [6] who have also discussed the motivations for considering these parsers for ASU. None of these parsers uses a stochastic grammar.

If island-driven parsers are used for generating partial interpretations of a spoken sentence, it is important to compute $\Pr(th)$, which is the probability that a SCFG generates sequences of words intermixed with *gaps* corresponding to portions of the acoustic signal that are still uninterpreted. Recent work provides a precise theoretical framework for this computation [3].

Many different cases involving islands and gaps have been examined; space considerations do not permit us to give here the lengthy formulas obtained for each of these cases. Instead, this paper will list the cases along with the worst-case time complexity of the computation of $\Pr(th)$ for each. Perhaps the most striking result of this work was the sharp division between the cases where one must compute the probability that a partial tree generates substrings of a sentence intermixed with a gap of **unknown** length, and the cases where the gap has a **known** length. The former computation appears to have an unacceptable time complexity; the latter computation is quite tractable. For this reason a later section considers ways in which one might estimate the length of a gap.

Definitions

A SCFG is a quadruple $G_s = (N, \Sigma, P, S)$, where N is a finite set of *nonterminal* symbols, Σ is a finite set of *terminal* symbols disjoint from N, P is a finite set of *productions* of the form $H \rightarrow \alpha$, $H \in N$, $\alpha \in (\Sigma \cup N)^*$, and $S \in N$ is a special symbol called *start symbol*. Each production is associated with a probability, indicated with $\Pr(H \rightarrow \alpha)$. If the grammar is *proper* the following relation holds:

$$\sum_{\alpha \in (\Sigma \cup N)^*} \Pr(H \rightarrow \alpha) = 1, \quad H \in N. \tag{1}$$

An SCFG G_s is in *Chomsky Normal Form* (CNF) if all productions in G_s are in one of the following forms:

$$H \rightarrow FG \quad H \rightarrow w, \quad H, F, G \in N, w \in \Sigma. \tag{2}$$

In the following we will always refer to SCFGs in CNF.

In the adopted formalism u, v and t represent strings of already recognized terminals; i, j and l are position indices;

p, q and r are shift indices; m indicates a (known) gap length and k, h are used as running indices. Furthermore, $x^{(m)}$ stands for a gap of unknown terminals with specified length m, while a gap of unknown terminals with unknown length is represented by $x^{(*)}$. Finally, Σ^* represents the set of all strings of finite length over Σ, while $\Sigma^m, m \geq 0$ is the set of all strings in Σ^* of length m.

The derivation of a string in G_s is usually represented as a parse (or derivation) tree, which shows the rules employed. It is also possible to associate with each derivation tree the probability that it was generated from a nonterminal symbol H by the grammar G_s. This probability is the product of the probabilities of all the rules employed in the derivation.

Given a string $x \in \Sigma^*$, the notation $H < x >$, $H \in N$, indicates the set of all trees with root H generated by G_s and spanning x. Therefore $\Pr(H < x >)$ is the sum of the probabilities of these subtrees, i.e. the probability that the string x has been generated by G_s starting from symbol H. We assume that the grammar G_s is *consistent*. This means that the following condition holds:

$$\sum_{x \in \Sigma^*} \Pr(S < x >) = 1. \tag{3}$$

From this hypothesis it follows that a similar condition holds for all nonterminals.

We are concerned with the computation of probabilities of strings involving islands. The assumed model of computation is the *Random Access Machine*, taken under the *uniform cost criterion* (see [1]). We will indicate with $|P|$ the size of set P, i.e. the number of productions in G_s. We will also write $f(x) = O(g(x))$ whenever there exist constants c, $\overline{x} > 0$ such that $f(x) > c\, g(x)$ for every $x > \overline{x}$. In the following section, we give the worst-case time complexity results we have derived.

Complexity Results

First, consider the computation of the probability that a given nonterminal H generates a tree whose yield is the string $ux^{(*)}vy^{(*)}$, where $u = w_i \ldots w_{i+p}$ and $v = w_j \ldots w_{j+q}$ are two already recognized substrings, while $x^{(*)}$ and $y^{(*)}$ represent two unspecified length gaps, i.e. two not yet specified strings of terminal symbols that can be generated in those positions by G_s. Such a probability will be indicated by $\Pr(H < ux^{(*)}vy^{(*)} >)$. For $H = S$, this probability gives the syntactical plausibility of the partial theory $ux^{(*)}vy^{(*)}$, which may be used for computing hypothesis scores in search of the most plausible interpretation of a spoken sentence. The asterisk indicates that nothing is known about gap x.

We have determined that calculation of such island probabilities with unknown gap length requires solving a rather huge non-linear system of $|N|(q+1)^2$ equations, q being the length of the island. If an approximate solution is of any interest, such a system can be rendered linear and can be solved by means of the computation of an $|N|(q+1)^2 \times |N|(q+1)^2$ inverse square matrix; this takes an $O(|N|^3 q^6)$ amount of time. For practical values of N and q the required computational effort seems unaffordable.

Tables 1, 2, and 3 list the remaining cases that have been examined, along with the worst-case time complexity of calculating each probability given a known SCFG G_s. Table 1 is self-explanatory. Table 2 deals with a problem of great practical interest - the computation of a theory that has been obtained from a previous theory by means of a single-word extension. In these cases the only calculation required concerns the new terms whose introduction is due to the added word. Table 3 shows the complexity of additional computation when not the *theory*, but the *gap*, is extended by one term. Since suffixes and prefixes are symmetric, the tables show only one of two symmetric cases (results still valid if strings are reversed).

The computations shown in Table 3 are particularly worth studying because we do not know *exactly* the number of words filling the gap but often know a probability distribution for this quantity; hence we have to take into account more than one value for the gap length.

Rows 3 and 5 in Table 3 show that a one-unit extension of a gap within a string costs a cubic amount of time (on top of work already done). If it is possible to get bounds on the number of (possible) words in a gap, this extra work will be repeated a fixed (in practical cases small) number of times.

Island-Driven Parsing Strategies

Given a method for scoring partial sentence interpretations in ASU systems, how can the method be utilized? This section discusses how the computations listed previously support island-driven bidirectional strategies for ASU.

In speech recognition and speech understanding tasks, partial theories are created and a strategy is used to select the most probable theory (theories) for growing.

The score of a theory *th* can be expressed as:

$$\Pr(ux^{(*)}vy^{(*)} \mid A) = \frac{\Pr(A \mid ux^{(*)}vy^{(*)}) \Pr(ux^{(*)}vy^{(*)})}{\Pr(A)}. \tag{4}$$

A parsing strategy can be considered that starts from left to right generating a sequence of word hypotheses u; subsequently, syntactic or semantic predictions generate a sequence v.

An upper bound for $\Pr(A|ux^{(*)}vy^{(*)})$ can be obtained by running the Viterbi algorithm using a model for u, followed

by a looped model of the lexicon (or the phonemes) for $x^{(*)}$, followed by a model for v and by a looped model for $y^{(*)}$.

Starting from *th*, a theory can grow by trying to fill the gap $x^{(*)}$ with a sequence of words. The hypotheses used for filling the gap may have one word, two words, three words, etc.. For each size of the gap an upper bound of the probability coming from the language model is $\Pr(ux^{(m)}vy^{(*)})$.

Reasonable assumptions about possible values of m can be obtained if suprasegmental acoustic cues such as energy contour descriptors are available. Based on a string A_g describing these features in the gap, it is possible to express the probability $\Pr(A_g|m)$ of observing A_g given a gap of m words as follows:

$$\Pr(A_g \mid m) = \sum_{s=0}^{\infty} \Pr(A_g \mid n_s = s, m) \Pr(n_s = s \mid m)$$
$$\cong \sum_{s=s_{min}}^{s_{max}} \Pr(A_g \mid n_s = s) \Pr(n_s = s \mid m) \quad (5)$$

where n_s indicates the number of syllables in the gap, and $\Pr(A_g \mid n_s = s)$ denotes the a-priori probability of observing A_g given that there are s syllables in the gap. It is reasonable to assume that this probability is a good approximation of the probability of observing A_g given that there are s syllables and m words in the gap. $\Pr(n_s = s \mid m)$ is the probability that a string of m words is made up of s syllables, and it can be estimated from a written text. The limits s_{min} and s_{max} are chosen in such a way that $\Pr(n_s = s \mid m) < \epsilon$ for $s < s_{min}$ and $s > s_{max}$, so that they depend on m and on the language model, but not on the input string and can be computed off-line.

Thanks to (5) it is possible to delimit practical values between which m can vary. Let m_1 and m_2 be the lowest and the highest value for m.

An upper bound for the probability of the language model relative to theory "th" can be expressed as:

$$U(ux^{(*)}vy^{(*)}) = \max_{m_1 \le m \le m_2} \Pr(ux^{(m)}vy^{(*)}) \quad (6)$$

We are mainly interested in ASU systems performing sentence interpretation in restricted domains. In this kind of task, non-syntactic information is usually available to predict words on the basis of previously obtained partial interpretations of the uttered sentence. Predicted words may be "islands" in the sense that they do not follow an existing partial theory in a strictly left-to-right manner. The acoustic evidence of these islands can be evaluated using word-spotting techniques. For these situations, *island-driven* parsers can be used. These parsers produce partial parses in which sequences of hypothesized words can be interleaved by gaps, making theories of the kind listed in the previous section (whose probabilities are calculated as described in [3]).

The same methods permit assessment of word candidates adjacent to an already recognized string - i.e., computation of the probability that the first (last) word of the gap x_1 (x_m) is a certain $a \in \Sigma$. This new word will extend the current theory. Normally, the system would select the word candidate(s) which maximize the prefix-string-with-gap probability of the theory augmented with it. Instead of computing these probabilities for all the elements in the dictionary, it is possible to restrict such an expensive process to the preterminal symbols (as in [8]).

The approach discussed here should be compared with standard lattice parsing techniques, where no restriction is imposed by the parser on the word search space (see, for example [4] and the discussion in [11]). Our framework accounts for bidirectional expansion of partial analyses; this improves the predictive capabilities of the system. In fact, bidirectional strategies can be used in restricting the syntactic search space for gaps surrounded by two partial analyses. This idea has been discussed without reference to stochastic grammars in [12] for the case of one word length gaps. We propose a generalization to m-length gaps and to cases where partial analyses do not represent entire parse trees but partial derivation trees.

A fair comparison between island-driven and left-to-right theory growing in stochastic parsing is not possible at present. In practice, island-driven parsers may remarkably accelerate the theory-growing process if island predictions are made by a look-ahead mechanism that leads to a correct partial theory with a limited number of competitors and if a limited number of predictions can be made for the words that can fill the gap.

HEURISTICS FOR IMPROVED LANGUAGE MODELING

The domains of discourse into which we might wish to introduce speech recognition systems vary widely. Often, the way in which human beings employ speech within a given domain has idiosyncracies which should be incorporated into the probabilistic language model, because they greatly increase its predictive power. In this section, we discuss two heuristics which may improve language modeling in specific situations.

Adding a Cache Component to a Standard Language Model

Our work on cache-based language modeling began with the simple observation that a given speaker or writer is likely

to use the same words repeatedly, and gave rise to a heuristic which greatly improved the performance of a standard probabilistic language model (the 3g-gram model). This heuristic is likely to be useful in any context where a speaker interacts with the speech recognition system for some length of time (the longer the interaction continues, the more accurate the system's estimate of the speaker's characteristic word use frequencies). Dictation systems are an obvious application, as is any interactive system in which the interaction is prolonged, or where the same people use the system repeatedly.

This section summarizes our work on cache components; see [9] for more details. Since the cache is superimposed on a standard language model (we used the 3g-gram model) we will begin with an overview of such models.

Consider the most straightforward kind of language modeling, where the task of the LM is to determine $Pr(th) = Pr(w_1...w_n)$. The trigram model [7] approximates this by setting $Pr(w_i|w_1...w_{i-1}) = Pr(w_i|w_{i-2}w_{i-1})$, which can be estimated from a training text. Thus we have

$$Pr(th) = Pr(w_1)Pr(w_2|w_1)Pr(w_3|w_1w2)...Pr(w_n|w_{n-2}w_{n-1}). \quad (7)$$

The 3g-gram model uses grammatical parts of speech (POS). Let $g(w_i) = g_i$ denote the POS of the word that appears at time i. Based on g_{i-1} and g_{i-2}, one part of the model gives the probability that g_i is a noun or a verb or an article, etc: $Pr(g_i = X) = Pr(g_i = X|g_{i-2}, g_{i-1})$. Another part gives the probability of a particular word if the POS is known. Both parts are estimated from frequencies in training texts. Thus, for a word W that has only one possible POS, $g(W)$, the probability $Pr(w_i = W)$ is estimated by

$$Pr(w_i = W|g_{i-2}, g_{i-1}) = Pr(W|g(W))Pr(g_i = g(W)|g_{i-2}, g_{i-1}). \quad (8)$$

Our contention was that while someone is speaking, a word used in the immediate past is very likely to be used again - much more likely than would be predicted by either of the models just described. We believed that these short-term word frequency fluctuations depend on the POS. Therefore, we used the 3g-gram model along with a **cache component** as the basis for a **combined model** which could weight the short-term cache component heavily for some POSs and not for others. The relative weights assigned to the cache and 3g-gram components within each POS category were obtained by maximum-likelihood estimation. To assess the improvement achieved by incorporating the cache component, we ran the combined model and a pure 3g-gram model (both trained on exactly the same data) on a text and compared the perplexities obtained.

The combined model gives a probability to each POS in the same way as the 3g-gram model. For a fixed POS, the probability of any word W which belongs to it is a weighted average of W's frequency in that POS category in the training text - the 3g-gram component - and W's frequency in the cache belonging to the POS category - the cache component. During the speech recognition task, the cache for a POS will contain the last N words which were guessed to have that POS (we set N to 200). If a word has occurred often in the recent past, it will occur many times in the cache for its POS.

Let $C_j(W, i)$ be the cache-based probability estimate for word W at time i for POS g_j. This is calculated from the frequency of W among the N=200 most recent words belonging to POS g_j. Our combined model estimates $Pr(w_i = W|g(W))$ by $k_{M,j}f(w_i = W|g(W)) + k_{C,j}C_j(W, i)$, where $k_{M,j} + k_{C,j} = 1$, instead of by $f(w_i = W|g(W))$ alone. The POS component $Pr(g_i = g(W)|g_{i-2}, g_{i-1})$ of the combined model was estimated as described in [5].

To train and test the pure 3g-gram model and the combined model, we utilized different portions of the LOB Corpus of British English: 100 sample texts drawn from this corpus form the testing text. The results exceeded our expectations. On the testing text, the pure 3g-gram model gave a perplexity of 332; the combined model gave a perplexity of 107. Thus, the incorporation of a cache component yielded a 3-fold improvement in perplexity.

The results confirm our hypothesis that recently-used words have a higher probability of occurrence than the 3g-gram model would predict, and that incorporating this knowledge into the LM via a cache component gives a significant improvement in performance.

The following are some ideas for extending this work:

- The weighting of the cache component could be made to depend on the number of words in the cache.
- The idea of tracking the speaker's recent behaviour could be extended to POSs - that is, recently employed POSs could be assigned higher probabilities.
- A word association matrix could be built, so that the occurrence of a word would increase the estimated probability of words that often co-occur with it.
- A cache component could be incorporated into a SCFG, where it would affect the probabilities of those productions which give rise to terminals.

Language Modeling and Semantics in Dialogue Systems

More recent work focuses on the special characteristics of dialogue. Given the context provided by the state of the discourse, there will be strong constraints on both syntax

and word choice which should be expressed *probabilistically* in the LM. Development of such an LM is one of our main current goals.

We have recently begun to consider the influence of discourse state on semantics. It has usually been tacitly assumed that almost all the words in an utterance must be correctly recognized for its meaning to be determined. This is true for isolated sentences, but it is seldom true during a dialogue. We wish to design a dialogue system capable of extracting semantic content even from distorted utterances, by using Bayesian criteria to decide between possible meanings. This work has links with the results for island-driven parsers, since meaningful word sequences often form islands within a user utterance. A rigorous quantitative theory linking these themes, together with the appropriate parsing algorithm, is under development.

REFERENCES

1. A.V.Aho J.E.Hopcroft and J.D.Ullman, "The Design Analysis of Computer Algorithms", Addison-Wesley Publishing Company, Reading, MA, 1974.
2. P.Brown F.Jelinek and R.L.Mercer, "Basic Method of Probabilistic Context Free Grammars", *Internal Report*, T.J.Watson Research Center, Yorktown Heights, NY 10598, 85 pages.
3. A.Corazza R.De Mori R.Gretter and G.Satta, "Computation of Probabilities for an Island-Driven Parser", *McGill University Technical Report*, No. SOCS 90.19, Jan. 1991.
4. Y.L.Chow and S.Roukos, "Speech Understanding Using a Unification Grammar", *Proc. IEEE International Conf. on Acoustic, Speech and Signal Processing*, 1989, Glasgow, Scotland, pp. 727-731.
5. A.M. Derouault and B. Merialdo, "Natural Language Modeling for Phoneme-to-Text Transcription", *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. PAMI-8, pp. 742-749, Nov. 1986.
6. E.P.Giachin and C.Rullent, "A Parallel Parser for Spoken Natural Language", *Proc. of Eleventh International Joint Conference on Artificial Intelligence*, 1989, Detroit, Michigan USA, pp.1537-1542.
7. F.Jelinek, "The Development of an Experimental Discrete Dictation Recognizer", *Proc.IEEE*, vol.73, n.11, Nov. 1985, pp. 1616-1624.
8. F.Jelinek, "Computation of the Probability of Initial Substring Generation by Stochastic Context Free Grammars", *Internal Report*, Continuous Speech Recognition Group, IBM Research, T.J.Watson Research Center, Yorktown Heights, NY 10598, 10 pages.
9. R.Kuhn and R.De Mori, "A Cache-Based Natural Language Model for Speech Recognition", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.12, n.6, June 1990, pp.570-583.
10. K.Lari and S.J.Young, "The Estimation of Stochastic Context-Free Grammars using the Inside-Outside Algorithm", *Computer Speech and Language*, vol.4, n.1, 1990, pp.35-56.
11. R.M.Moore F.Pereira and H.Murveit, "Integrating Speech and Natural Language Processing", *Proceedings of the Speech and Natural Language Workshop*, 1989, Philadelphia, Pennsylvania, pp.243-247.
12. O.Stock R.Falcone and P.Insinnamo, "Bidirectional Chart: A Potential Technique for Parsing Spoken Natural Language Sentences", *Computer speech and Language*, vol.3, n.3, 1989, pp.219-237.
13. W.A.Woods, "Optimal Search Strategies for Speech Understanding Control", *Artificial Intelligence*, vol.18, n.3, 1981, pp.295-326.

Computed probability and its time complexity

gap probabilities

$1.\{\Pr(H < x^{(k)} >) \mid H \in N, 1 \leq k \leq m\}$
$-> O(|P|m^2)$

inside probabilities

$2.\{\Pr(H < w_i \dots w_{i+p} >) \mid H \in N\}$
$-> O(|P|p^3)$

prefix/suffix-string probabilities

$3.\{\Pr(H < w_i \dots w_{i+p} x^{(*)} >) \mid H \in N\}$
$-> O(|P|p^3)$
$4.\{\Pr(H < w_i \dots w_{i+p} x^{(m)} >) \mid H \in N\}$
$-> O(|P| \max\{p^3, m^2 p\})$

gap-in-string probabilities

$5.\{\Pr(H < w_i \dots w_{i+p} x^{(*)} w_j \dots w_{j+q} >) \mid H \in N\}$
$-> O(|P| \max\{p^3, q^3\})$
$6.\{\Pr(H < w_i \dots w_{i+p} x^{(m)} w_j \dots w_{j+q} >) \mid H \in N\}$
$-> O(|P| \max\{p^3, q^3, m^2 p\})$

island probabilities

$7.\{\Pr(H < x^{(m)} w_j \dots w_{j+q} y^{(*)} >) \mid H \in N\}$
$-> O(|P| \max\{q^3, m^2 q\})$

prefix-string-with-gap probabilities

$8.\{\Pr(H < w_i \dots w_{i+p} x^{(m)} w_j \dots w_{j+q} y^{(*)} >) \mid H \in N\}$
$-> O(|P| \max\{p^3, q^3, pm^2, qm^2\})$

Table 1: *Worst-case time complexity for the computation of bounded and unbounded gap length probabilities.*

Prob. of theory extended by word a
inside probabilities
1.$\{\Pr(H < w_i \dots w_{i+p} a >) \mid H \in N\}$ $-> O(\lvert P\rvert p^2)$
prefix-string probabilities
2.$\{\Pr(H < w_i \dots w_{i+p} a x^{(*)} >) \mid H \in N\}$ $-> O(\lvert P\rvert p^2)$
3.$\{\Pr(H < a w_i \dots w_{i+p} x^{(m)} >) \mid H \in N\}$ $-> O(\lvert P\rvert \max\{p^2, m^2\})$
4.$\{\Pr(H < w_i \dots w_{i+p} a x^{(m-1)} >) \mid H \in N\}$ $-> O(\lvert P\rvert \max\{p^2 m, p m^2\})$
gap-in-string probabilities
5.$\{\Pr(H < w_i \dots w_{i+p} x^{(*)} w_j \dots w_{j+q} a >) \mid H \in N\}$ $-> O(\lvert P\rvert \max\{p^2, q^2\})$
6.$\{\Pr(H < w_i \dots w_{i+p} x^{(*)} a w_j \dots w_{j+q} >) \mid H \in N\}$ $-> O(\lvert P\rvert \max\{p^2 q, p q^2\})$
7.$\{\Pr(H < w_i \dots w_{i+p} x^{(m)} w_j \dots w_{j+q} a >) \mid H \in N\}$ $-> O(\lvert P\rvert \max\{p^2, q^2, m^2, p(m+q)\})$
8.$\{\Pr(H < w_i \dots w_{i+p} x^{(m)} a w_j \dots w_{j+q} >) \mid H \in N\}$ $-> O(\lvert P\rvert \max\{p^2 q, p q^2, q^2 m, q m^2\})$
9.$\{\Pr(H < w_i \dots w_{i+p} x^{(*)} a >) \mid H \in N\}$ $-> O(\lvert P\rvert p^2)$
10.$\{\Pr(H < w_i \dots w_{i+p} x^{(m)} a >) \mid H \in N\}$ $-> O(\lvert P\rvert \max\{p^2, m^2\})$
island probabilities
11.$\{\Pr(H < x^{(m)} w_j \dots w_{j+q} a y^{(*)} >) \mid H \in N\}$ $-> O(\lvert P\rvert \max\{q^2, m^2\})$
12.$\{\Pr(H < x^{(m)} a w_j \dots w_{j+q} y^{(*)} >) \mid H \in N\}$ $-> O(\lvert P\rvert \max\{m^2 q, m q^2\})$
prefix-string-with-gap probabilities
13.$\{\Pr(H < w_i \dots w_{i+p} a x^{(m-1)} w_j \dots w_{j+q} y^{(*)} >) \mid H \in N\}$ $-> O(\lvert P\rvert \max\{p^2 q, p q^2, p^2 m, p m^2\})$
14.$\{\Pr(H < w_i \dots w_{i+p} x^{(m-1)} a w_j \dots w_{j+q} y^{(*)} >) \mid H \in N\}$ $-> O(\lvert P\rvert \max\{p^2 q, p q^2, p^2 m, p m^2\})$
15.$\{\Pr(H < w_i \dots w_{i+p} x^{(m)} w_j \dots w_{j+q} a y^{(*)} >) \mid H \in N\}$ $-> O(\lvert P\rvert \max\{p^2, q^2, m^2, (m+q)p\})$
16.$\{\Pr(H < w_i \dots w_{i+p} x^{(m)} a y^{(*)} >) \mid H \in N\}$ $-> O(\lvert P\rvert \max\{m^2, p^2\})$

Table 2: *Worst-case time complexity for the computation of probabilities of theories extended by the addition of a single word* $a \in \Sigma$.

Prob. of one-unit extension of known-length gap
gap probabilities
1.$\{\Pr(H < x_1^{(m)} x_2^{(1)} >) \mid H \in N\}$ $-> O(\lvert P\rvert m)$
prefix-string probabilities
2.$\{\Pr(H < w_i \dots w_{i+p} x_1^{(1)} x_2^{(m)} >) \mid H \in N\}$ $-> O(\lvert P\rvert \max\{p^2, pm\})$
gap-in-string probabilities
3.$\{\Pr(H < w_i \dots w_{i+p} x_1^{(m)} x_2^{(1)} w_j \dots w_{j+q} >) \mid H \in N\}$ $-> O(\lvert P\rvert \max\{p^2 q, q p^2, pnq\})$
island probabilities
4.$\{\Pr(H < x_1^{(m)} x_2^{(1)} w_j \dots w_{j+q} y^{(*)} >) \mid H \in N\}$ $-> O(\lvert P\rvert \max\{q^2, qm\})$
prefix-string-with-gap probabilities
5.$\{\Pr(H < w_i \dots w_{i+p} x_1^{(m)} x_2^{(1)} w_j \dots w_{j+m} y^{(*)} >) \mid H \in N\}$ $-> O(\lvert P\rvert \max\{p^2 q, p q^2, (q+m)p\})$

Table 3: *Worst-case time complexity for the computation of probabilities of theories extended by means of incrementing by one unit the known length gap.*