Weakly Restricted Stochastic Grammars

Rieks op den Akker and Hugo ter Doest

University of Twente, Department of Computer Science
P.O.Box 217, 7500 AE Enschede, The Netherlands

Keywords: stochastic languages, grammars, grammar inference.

Abstract

A new type of stochastic grammars is introduced for investigation: weakly restricted stochastic grammars. In this paper we will concentrate on the consistency problem. To find conditions for stochastic grammars to be consistent, the theory of multitype Galton-Watson branching processes and generating functions is of central importance. The unrestricted stochastic grammar formalism generates the same class of languages as the weakly restricted formalism. The inside-outside algorithm is adapted for use with weakly restricted grammars.

1 Introduction

If we consider a natural languages as a structure modelled by a formal grammar we do not consider it any more as a language that is used. Formal (context-free) grammars are often advocated as a model for the "linguistic competence" of an ideal natural language user. It is also noticed that this mathematical concept is far from a sufficient model for describing all aspects of the language. What cannot be expressed by this model is the fact that some sentences or phrases are more likeley to occur than others. This notion of occurrence refers to the use of language and therefore considering this kind of statistical knowledge about language has to do with the pragmatics of language laid down in a corpus of the language. With a particular context of use in mind a syntactically ambiguous sentence will often have a most likely meaning and hence a most likely analysis. Some of the shortcomings of the pure (context-free) grammar model can maybe be solved by stochastic grammars, a model that makes it possible to incorporate certain statistical facts about the language use into a model of the possible structures of sentences as we conceive them from a mathematical, formal, point of view. Natural languages are now seen as stochastic; a user of a language as a stochastic source producing sentences. A stochastic language over some alphabet Σ is simply a formal language L over Σ together with a probability function ϕ assigning to each string x in the language a real number $\phi(x)$ in $[0,1]$. Since $\phi(x)$ is interpreted as the chance that the event x, or the event that a language-source produces x, will occur, it will be clear that the sum of $\phi(x)$ where x ranges over all possible sentences is equal to one. The stochastic language is called context-free if the language L is context-free.

The usual grammatical model for a stochastic context-free language is a context-free grammar together with a probability function f that assigns a real number in $[0,1]$ to each of the productions of the grammar. The meaning of this function is the following. A step in a derivation of a sentential form, in which a nonterminal A is rewritten using production p has chance $f(p)$ to occur, independent of which A is rewritten in the sentential form and independent of the history of the proces that produced the sentential form. The probability of a derivation(-tree) is the product of the probabilities of the derivation steps that produces the tree. The probability of a sentence generated by the grammar is the sum of the probabilities of all the trees of a sentence. So given a stochastic grammar we can compute the probabilities of all its sentences. The distribution language generated by a stochastic grammar G, $DL(G)$, is defined as the set of all derivation trees with their probabilities. The stochastic language generated by a stochastic grammar G, $SL(G)$, is defined as the set of all sentences generated by the grammar with their probabilities. A stochastic grammar G is an adequate model of a language L if on its basis we can correctly compute the probabilities of the sentences in the language L. Of course this assumes a statistical analysis of a language corpus. A stochastic grammar that generates a stochastic language is called consistent.

Definition 1.1 *A stochastic grammar G is called* consistent *if for the probability measure p induced by G onto the language generated by its underlying grammar:*

$$\sum_{x \in L(G)} p(x) = 1$$

Otherwise the grammar is called inconsistent.

Not all stochastic grammars generate a stochastic language. Even proper, and reduced grammars [1] do not

[1] A grammed is called *proper* if for all nonterminals A, the sum of the probabilities assigned to the rules for A is 1. A grammar is called *reduced* if all nonterminals are reachable and can produce a terminal string.

necessarily generate a stochastic language. This is illustrated in the following example. [2]

Example 1.1 Consider the stochastic grammar G with nonterminal set $V_N = \{S\}$, terminal set $V_T = \{a\}$. The productions with their probabilities are given by:

$$S \xrightarrow{q} S\ S$$
$$S \xrightarrow{1-q} a$$

Following the technique presented in [2] we find that the production generating function is given by $g_1(s_1) = qs_1^2 + 1 - q$, and that the first moment matrix E is given by $[2q]$. We can conclude that the grammar is consistent if and only if $q \leq 1/2$. For details we refer to [5]. Notice that all the different trees of string a^n have the same probability. Hence, they cannot be distinguished according to their probabilities. □

It has been noticed that the usual model of a stochastic grammar as presented above, and which we from now on call the *unrestricted stochastic grammar model*, has some disadvantages for modelling "real" languages. In this paper we present a more adequate model, the weakly restricted stochastic grammar model. We give necessary and sufficient conditions to test in an efficient way whether such a grammar defines a stochastic language. Moreover, we will show that these grammars can be transformed into an equivalent model of the usual type. The nice thing about the new model is that it models "context-dependent" probabilities of production-rules directly in terms of the grammar specification of the language and not in terms of some particular implementation of the grammar as a parser. The latter is done by Briscoe and Carroll [3] by assigning probabilities to the transitions of the LR-parser constructed for the grammar. In section 2 weakly restricted grammars are introduced, in section 3 conditions for their consistency are investigated; in section 4 it is proven that weakly restricted grammars and unrestricted grammars generate the same class of stochastic languages and section 5 presents the inside-outside algorithm for weakly restricted grammars.

2 Weakly Restricted Stochastic Grammars

To add context-sensitivity to the assignment of probabilities to the application of production rules, we take into account (and distinguish) the occurrences of the nonterminals. Then, for each nonterminal occurrence distinct probabilities can be given for the production rules that can be used to rewrite the nonterminal. This way of assigning probabilities to the application of production rules seems unknown in literature, although we found some other formalisms that were designed to add context-sensitivity to the assignment of probabilities. For instance, the definition of stochastic grammars by Salomaa in [8] is somewhat different from the definition we gave in our introduction: the probability of a production to be applied is here dependent on the production that was last applied.

To escape the bootstrap problem (when a derivations is started, there is no last applied production) an initial stochastic vector is added to the grammar. Weakly restricted stochastic grammars are introduced in [1]. In the following definition C_{A_i} denotes the set of productions for A_i and $R(A_i)$ denotes the number of right-hand side occurrences of nonterminal A_i.

Definition 2.1 *A weakly restricted stochastic grammar G_w is a pair (G_c, Δ), where $G_c = (V_N, V_T, P, S)$ is a context-free grammar and Δ is a set of functions*

$$\Delta = \{p_i | A_i \in V_N\}$$

where, if $j \in 1 \ldots R(A_i)$ and $k \in 1 \ldots |C_{A_i}|$, $p_i(j,k) = p_{ijk} \in [0,1]$ The set of productions P contains exactly one production for start symbol S.

In words, p_{ijk} stands for the probability that the k-th production with left-hand side A_i is used for rewriting the j-th right-hand side-occurrence of nonterminal A_i. The usefullness of this context-dependency can be seen immediately from the following unrestricted stochastic grammar, which is taken (in part) from the example grammar in [3] (p. 29):

$$S \xrightarrow{1.0} NP\ VP$$
$$VP \xrightarrow{0.4} Vt\ NP$$
$$VP \xrightarrow{0.6} Vi$$
$$NP \xrightarrow{0.4} ProNP$$
$$NP \xrightarrow{0.3} Det\ N$$
$$NP \xrightarrow{0.3} NP\ PP$$

Unrestricted stochastic grammars cannot model context dependent use of productions. For example, an NP is more likely to be expanded as a pronoun in subject position than elsewhere. Exactly this dependence on where a nonterminal was introduced can be modeled by using a weakly restricted stochastic grammar. Since in a weakly restricted stochastic grammar the probabilities of applying a production are dependent on the particular occurrence of a nonterminal in the right-hand side of a production, it is useful to require that there is only one start production.

The characteristic grammar of a weakly restricted grammar is the underlying context free grammar. The next step is to compute probabilities for strings with respect to weakly restricted stochastic grammars. For this purpose a tree is written in terms of its subtrees (trees with a nonterminal as root) as $q[t_{i_1 j_1}, t_{i_2 j_2}, \ldots, t_{i_{n(q)} j_{n(q)}}]$, in which q is a production, $n(q)$ is the number of nonterminals in the right-hand side of q and t_{ij} denotes a (sub)tree with the j-th

[2] Although we found in [7] by Jelinek and Lafferty the (false) statement that a stochastic grammar is consistent if and only if it is proper, given that the underlying grammar is reduced. The example gives a clear counter example of their statement.

occurrence of nonterminal A_i at its root. A tree for which $n(q) = 0$ is written as $[]$.

Definition 2.2 *The probability of a derivation tree t with respect to a weakly restricted stochastic grammar is defined recursively a*

$$P_w([]) = 1$$
$$P_w(q[t_{i_1 j_1}, t_{i_2 j_2}, \ldots, t_{i_{n(q)} j_{n(q)}}]) = \prod_{m=1}^{n(q)} p_{i_m j_m k_m} P_w(t_{i_m j_m})$$

where $1 \leq k_m \leq |C_{A_{i_m}}|$.

The probability of a string is defined as the sum of the probabilities of all distinct derivation trees that yield this string.

Definition 2.3 *The probability of a string x in $L(G_c)$ is defined as*

$$p_w(x) = \sum \{P_w(t) | (x, P_w(t)) \in DL(G)\}$$

The distribution language $DL(G_w)$ and stochastic language $SL(G_w)$ of a weakly restricted grammar (G_c, Δ) are defined analoguous to the distribution language and stochastic language of an unrestricted grammar.

3 Consistency

In this section consistency of weakly restricted stochastic grammars will be considered. The theory of multitype branching processes will be used to come to a similar theorem as is given in [2] for unrestricted stochastic grammars.

Definition 3.1 *For the j-th occurrence of nonterminal $A_i \in V_N$ the* production generating function *for weakly restricted stochastic grammars is defined as:*

$$g_{ij}(s_{1,1}, \ldots, s_{k,R(A_k)}) = \sum_{u=1}^{|C_{A_i}|} p_{iju} \prod_{m=1}^{k} \prod_{n=1}^{R(A_m)} s_{mn}^{r_{mn}(u)}$$

where $r_{mn}(k)$ is 1 if nonterminal-occurrence A_{mn} appears in the right-hand side of the k-th production rule with nonterminal A_i as left-hand side and 0 otherwise.

Note that for each right-hand side nonterminal occurrence a dummy-variable is introduced: s_{ij} corresponds to the j-th occurrence of nonterminal A_i. A special variable is $s_{1,1}$: it corresponds to the start symbol which is the right-hand side of the start production $s \in P$ of the form $Z \rightarrow S$. The generating function for nonterminal occurrence A_{ij} entails for each production for A_i a term. If g_{ij} has a term of the form

$$\alpha s_{i_1} s_{i_2} \ldots s_{i_n},$$

then we know that it corresponds to a production for A_i of the form

$$A_i \rightarrow x_{i_1} A_{i_1} x_{i_2} A_{i_2} \ldots x_{i_n} A_{i_n} x_{i_{n+1}}$$

where the $x_{i_j} \in V_T^*$. The production has, if it is used for rewriting occurrence A_{ij}, probability α of being applied. In Example 3 it will be illustrated how the terms of the generating functions correspond to the productions of the grammar.

Theorem 3.1 *Let $A_{ij} \Rightarrow \alpha$; thus the j-th occurrence of nonterminal A_i is rewritten using exactly one production. The probability that α contains the n-th occurrence of nonterminal A_m is given by*

$$c_{ijmn} = \left. \frac{\partial g_{ij}(s_{1,1}, \ldots, s_{k,R(A_k)})}{\partial s_{mn}} \right|_{s_{1,1}, \ldots, s_{k,R(A_k)} = 1}$$

Proof In general the generating function can be written as

$$g_{ij}(s_{1,1}, \ldots, s_{k,R(A_k)}) = g'_{ij}(s_{1,1}, \ldots, s_{k,R(A_k)}) + c_{ij}$$

where $g'_{ij}(s_{1,1}, \ldots, s_{k,R(A_k)})$ only contains terms dependent on $s_{1,1}, \ldots, s_{k,R(A_k)}$ and where c_{ij} is a constant term. The terms dependent on $s_{1,1}, \ldots, s_{k,R(A_k)}$ come from productions for A_i that contain nonterminals in their right-hand sides and the constant terms from productions for A_i that only contain terminals in their right-hand sides. When partial derivatives are taken from g_{ij} we can just as well consider g'_{ij}, since the constant term will become zero. We know that the terms in g'_{ij} do not contain any powers higher than 1 of the variables in it. This leads us to the insight that taking the mn-th partial derivative of g_{ij} results in at most one term consisting of the form $p_{ij*} f(s_{1,1}, \ldots, s_{k,R(A_k)})$ where f does not depend on s_{mn} and p_{ij*} is one of the probabilities resulting from applying p_i to j and some k in $1 \ldots |C_{A_i}|$. If we substitute 1 for all remaining variables in the partial derivative we find as value for c_{ijmn} the probability that the j-th occurrence of nonterminal A_i is rewritten by the production that contains in its right-hand side nonterminal occurrence A_{mn}. □

The first-moment matrix for weakly restricted grammars is defined just like the first-moment matrix for unrestricted grammars:

Definition 3.2 *The* first-moment matrix E *associated with the weakly restricted grammar G is*

$$E = [e_{ijmn}]$$

where $1 \leq i, m \leq |V_N|$ and $1 \leq j, n \leq R(A_i)$.

We order the set of eigenvalues of the first-moment matrix from the largest one to the smallest, such that ρ_1 presents the maximum.

Theorem 3.2 *A proper weakly restricted grammar is consistent if $\rho_1 < 1$ and is not consistent if $\rho_1 > 1$*

The proof of this theorem is analoguous to the proof of the related theorem in [2] and we will not treat it here (see [5] for a proof).

Example 3.1 Consider the weakly restricted stochastic grammar (G_c, Δ) where $G_c = (V_N, V_T, P, Z) = (\{Z, S\}, \{a\}, P, Z)$ and P and Δ are as follows:

$$\begin{array}{ll} Z \rightarrow S & (p, 1-p) \\ S \rightarrow S\ S & (q, 1-q)(r, 1-r) \\ S \rightarrow a & \end{array}$$

For a reason at the of the example to become clear, we assume that $p \neq 0$. The production generating functions are given by

$$\begin{aligned} g_{11}(s_{11}, s_{12}, s_{13}) &= ps_{12}s_{13} + 1 - p \\ g_{12}(s_{11}, s_{12}, s_{13}) &= qs_{12}s_{13} + 1 - q \\ g_{13}(s_{11}, s_{12}, s_{13}) &= rs_{12}s_{13} + 1 - r \end{aligned}$$

The first-moment matrix E is given by

$$\begin{bmatrix} 0 & p & p \\ 0 & q & q \\ 0 & r & r \end{bmatrix}$$

The characteristic equation is given by $\phi(x) = x((x-q)(q-r)-qr) = x^2(x-(q+r)) = 0$. Thus, the eigenvalues of the matrix are 0 and $x = q + r$. According to Theorem 3.2 the grammar is consistent if $q+r < 1$ and inconsistent if $q + r > 1$. If $q + r = 1$ the theorem does not decide the consistency of the grammar. From the characteristic equation it follows that the value of p does not influence the consistency of the grammar. However, looking at the grammar we find that it is consistent if $p = 0$, regardless of probabilities q and r. Therefore, before Theorem 3.2 can be used for checking the consistency of the grammar, the grammar must be stripped of productions having for each nonterminal occurrence probability zero of being applied. □

Definition 3.3 *A* final class C *of nonterminal occurrences is a subset of the set of all nonterminal occurrences having the property that any occurrence in* C *has probability* 1 *of producing, when rewritten using one production rule, exactly one occurrence also in* C.

Theorem 3.3 *A weakly restricted stochastic grammar is consistent if and only if* $\rho_1 \leq 1$ *and there are no final classes.*

For the proof of Theorem 3.3 we refer to [5]. Applying this theorem to the example learns us that if $q + r = 1$, the grammar is consistent if and only if there is no final class of nonterminals. Looking at the grammar we see that there is a final class of occurrences if $q = 1$ or $r = 1$ (or both); the final classes then are $\{S_2\}$,$\{S_3\}$ and $\{S_2, S_3\}$, respectively; if in addition $p = 1$, then the final classes are $\{S_1, S_2\}$,$\{S_1, S_3\}$ and $\{S_1, S_2, S_3\}$, respectively. Hence, the grammar is consistent if and only if $q + r \leq 1 \wedge q \neq 1 \wedge r \neq 1$. Notice that if $q \neq r$ then all trees of a^n have different probabilities.

4 Equivalence

In this section we will show that a weakly restricted stochastic grammar can be transformed into an equivalent unrestricted grammar. We define two grammars G and H to be equivalent if $DL(G) = DL(H)$.

The transformation is performed as follows. With each nonterminal occurrence A_{ij} in the right-hand side of a production rule associate a new unique nonterminal Aij; for each new nonterminal Aij copy the set of production rules with nonterminal A_i as left-hand side, replace the left-hand sides with Aij and replace in the right-hand sides each nonterminal with its new (associated) nonterminal; assign probability p_{ijk} to the k-th production rule with left-hand side A_{ij}. We formalized this in the following algorithm.

Algorithm 4.1

1 Associate with the j-th occurrence of nonterminal A_i in the right-hand sides of the production rules a (new) unique nonterminal A_{ij} (clearly $j \in 1, \ldots, R(A_i)$). The set of nonterminals for the rewritten grammar G' is denoted by V'_N and is the set of associated nonterminals plus the start symbol S from the weakly restricted grammar G.

2 This step is given in pseudo-pascal:

for $i := 1$ **to** $|V_N|$ **do**
  for $j := 1$ **to** $R(A_i)$ **do**
    $P' := P' \cup C_{A_i}(j)$
  od
od

where $C_{A_i}(j)$ is the set of productions C_{A_i} with left-hand sides A_i replaced by A_{ij} and the nonterminals in the right-hand sides of the production rules replaced by their associated nonterminals.

3 The probabilities to be assigned to the production rules in $C_{A_i}(j)$ are deduced from the $p_{ij} = (p_{ij1}, \ldots, p_{ij|C_{A_i}|})$: the k-th production rule in $C_{A_i}(j)$ is assigned probability p_{ijk}.

□

Theorem 4.1 *For every weakly restricted stochastic grammar there is an unrestricted stochastic grammar which is distributively equivalent.*

Proof We can prove the theorem by proving that the algorithm finds for every weakly restricted grammar an unrestricted grammar that is distributively equivalent. From the algorithm it immediately follows that the languages (without the probabilities) generated by the weakly restricted grammar and the unrestricted grammar generated by the algorithm are equal. The production rules introduced by the algorithm in the unrestricted grammar cannot generate any other strings than the string generated by

the weakly restricted grammar. Also it can be seen from the algorithm that the unrestricted grammar associates the same probabilities with its strings as the unrestricted grammar. Hence, the theorem holds. □

A corollary of this theorem is that for each weakly restricted grammar there exists an unrestricted grammar that is *stochastically equivalent.*

The time-complexity of the algorithm can easily be found. We observe that, if we denote the number of nonterminals in the weakly restricted grammar by k, each step can be done in in $O(k)$ steps. Then the total time complexity is $O(k)$. We define the size of a grammar to be the product of the number of nonterminals and the number of productions. The size of the newly created grammar can be found to be polynomial in the size of the weakly restricted grammar.

5 Inference

The inside-outside algorithm is originally a reestimation procedure for the rule probabilities of an unrestricted stochastic grammar in Chomksy Normal Form (CNF) [4]. It takes as input an initial unrestricted stochastic grammar G in CNF and a sample set E of strings and it iteratively recstimates rule probabilities to maximize the probability that the grammar would produce the sample set.

The basic idea of the inside-outside algorithm is to use the current rule probabilities to estimate from the sample set the expected frequencies of certain derivation steps, and then compute new rule probability estimates as appropriate frequency rates. Therefore, each iteration of the algorithm starts by calculating the *inside* and *outside* probabilities for all strings in the sample set. These probabilities are in fact probability functions which have as arguments a string w from the sample set, indexes which indicate what substring of w is to be considered, and an occurrence of a nonterminal, say A. With these arguments, the inside probability now is the probability that the occurrence of A derives the substring of w; the outside probability is the probability that the occurrence of nonterminal A appears in the intermediate string of some derivation of string w.

In what follows, we will take V_N,V_T as fixed $n = |V_N|$, $t = |V_T|$, and assume that $V_N = \{Z = A_0, S = A_1, A_2, \ldots, A_n\}$ and $V_T = \{a_1, \ldots, a_t\}$. By definition it is required that the grammar has one production for start symbol Z: $Z \rightarrow S$. Parallel to the definition of generating functions for weakly restricted grammars, we have to distinguish all nonterminal occurrences in right-hand sides of productions; we remind that the probability of each production depends on the particular nonterminal occurrence to be rewritten. The inside and outside probabilities now have to be specified for each nonterminal occurrence seperately. As already stated in the introduction, the inside-outside algorithm is designed only for context-free grammars in CNF. Using this fact we can simplify the way nonterminal occurrences are indexed: $A_{q(p.r)}$ ($A_{r(pq.)}$) denotes the occurrence of A_q (A_r) in the production $A_p \rightarrow A_q A_r$; for this production also the notation (pqr) is used and for the production $A_p \rightarrow a_q$ (pq). Similarly the probability of occurrence $A_{q(p.r)}$ to be rewritten using rule (qst) is denotes by $p_{q(p.r)(qst)}$. For the start production a special provision has to be taken: the nonterminal occurrence in its right-hand side is denoted by $A_{1(0..)}$. A stochastic grammar in CNF over these sets can then be specified by

$$\sum_i R(A_i)|P|$$

probabilities. Since we require stochastic grammars to be proper, we know that for $p, q, r = 1, \ldots, n$

$$\sum_{s,t} p_{q(p.r)(qst)} + \sum_s p_{q(p.r)(qs)} = 1$$

If we want to use the inside-outside algorithm for grammar inference, then the grammar probabilities have to meet the above condition in order for the reestimation to make sense.

If string $w = w_1 w_2 \ldots w_{|w|}$, then ${}_i w_j$, $0 \leq i < j < |w|$ denotes the substring $w_{i+1} \ldots w_j$. The inside probability $I^w_{p(q.r)}(i,j)$ estimates the likelihood that occurrence $A_{p(q.r)}$ derives ${}_i w_j$, while the outside probability $O^w_{p(q.r)}(i,j)$ estimates the likelihood of deriving ${}_0 w_i A_{p(q.r)j} w_{|w|}$ from the start symbol S. The inside-probability for string w and nonterminal occurrence $A_{p(q.r)}$ is defined by the recurrent relation

$$I^w_{p(q.r)}(i-1,i) = p_{p(q.r)(ps)}, \text{ where } a_s = w_i$$
$$I^w_{p(q.r)}(i,k) = \sum_{s,t} \sum_{i<j<k} p_{p(q.r)(pst)} I^w_{s(p.t)}(i,j) I^w_{t(ps.)(j,k)}$$

Similarly, the outside probabilities for shorter spans of w can be computed from the inside probabilities and the outside probabilities for longer spans by the following recurrence:

$$O^w_{p(q.r)}(0,|w|) = 1, \text{ if } q = 1$$
$$O^w_{p(q.r)}(0,|w|) = 0, \text{ otherwise}$$
$$O^w_{p(q.r)}(i,k) = \sum_{s,t} \sum_{j=0}^{i-1} O^w_{q(s.t)}(j,k) I^w_{r(qp.)}(j,i) p_{q(s.t)(qpr)}$$

The second equation above is somewhat simpler than the corresponding one for unrestricted stochastic grammars, because the occurrence $A_{p(q.r)}$ for which the outside probability $O^w_{p(q.r)}(i,k)$ is computed specifies the production used for creating it and consequently the probability for $A_{p(q.r)}$ to generate ${}_0 w_i A_{p(q.r)j} w_{|w|}$ is the sum of much less possibilities.

Once the inside and outside probabilities are computed for each string in the sample set E, the reestimated probability of binary rules, $\hat{p}_{p(q.r)(pst)}$, and the

reestimated probability of unary rules, $\hat{p}_{q(p.r)(qr)}$, are computed using the following reestimation formulae:

$$\hat{p}_{p(q.r)(pst)} = \frac{\sum_{w \in E} \frac{1}{P^w} \sum_{0 \le i < j < k \le |w|} \left[\begin{array}{c} p_{p(q.r)(pst)} I^w_{s(q.t)}(i,j) \\ \times \\ I^w_{t(ps.)}(j,k) O^w_{p(q.r)}(i,k) \end{array} \right]}{\sum_{w \in E} P^w_{p(q.r)} / P^w}$$

$$\hat{p}_{p(q.r)(ps)} = \frac{\sum_{w \in E} \frac{1}{P^w} \sum_{1 \le i \le |x|, w_i = a_s} p_{p(q.r)(ps)} O^w_{p(q.r)}(i-1,i)}{\sum_{w \in E} P^w_{p(q.r)} / P^w}$$

where P^w is the probability assigned by the current model to string w

$$P^w = I^w_{1(0..)}(0, |w|)$$

and P^w_p is the probability assigned by the current model to the set of derivations involving some instance of A_p

$$P^w_{p(q.r)} = \sum_{0 \le i < j \le |w|} I^w_{p(q.r)}(i,j) O^w_{p(q.r)}(i,j)$$

The denominator of the estimates $\hat{p}_{p(q.r)(pst)}$ and $\hat{p}_{p(q.r)(ps)}$ estimates the probability that a derivation of a string $w \in E$ will involve at least one expansion of the nonterminal occurrence $A_{p(q.r)}$. The numerator of $\hat{p}_{p(q.r)(pst)}$ estimates the probability that a derivation of a string $w \in E$ will involve rule $A_p \rightarrow A_q A_r$, while the numerator of $\hat{p}_{p(q.r)(ps)}$ estimates the probability that a derivation of a string $w \in E$ will rewrite A_p to a_s. Thus $\hat{p}_{p(q.r)(pst)}$ estimates the probability that a rewrite of $A_{p(q.r)}$ in a string from E will use rule $A_p \rightarrow A_s A_t$, and $\hat{p}_{p(q.r)(ps)}$ estimates the probability that occurrence $A_{p(q.r)}$ in a string from E will be rewritten to a_s. Clearly, these are the best current estimates for the binary and unary rule probabilities. The process is then repeated with the reestimated probabilities until the increase in the estimated probability of the sample set given the model becomes negligible. We presented the inside, outside and (estimated) production probabilities only for the nonterminal occurrences of the form $A_{p(q.r)}$; for occurrences $A_{p(qr.)}$ these can simply be found by adapting the equations we have given for them.

The reestimation algorithm can be used both to refine the current estimated probabilities of a stochastic grammar and to infer a stochastic grammar from scratch. The former application can be said to be incremental. In the latter case, the initial weakly restricted grammar for the inside-outside algorithm consists of all possible CNF rules over the given sets V_N of nonterminals and V_T of terminals, with suitable nonzero probabilities assigned to the nonterminal occurrences.

6 Conclusions

In this paper we have investigated consistency of weakly restricted stochastic grammars and presented an adapted version of the inside-outside algorithm. Other issues concerning stochastic grammars and especially weakly restricted grammars that are being investigated at the moment are stochastic grammatical inference and parsing using weakly restricted grammars. By stochastic grammatical inference we mean grammatical inference whereby the production probabilities are computed simultaneously. Consistency of stochastic grammars and stochastic inference will be treated in full in the master thesis of H.W.L. ter Doest, which is to appear in 1994 [5].

Acknowledgement We are grateful to Jorma Tarhio, presently at the University of Berkeley, California, for stimulating discussions.

References

[1] R. op den Akker. *Stochastic Grammars: theory and applications.* University of Twente, Department of Computer Science, Memoranda Informatica 93-19, 1993.

[2] T.L. Booth, R.A. Thompson. Applying Probability Measures to Abstract Languages. In: *IEEE Transactions on Computers*Vol. C-22, No. 5, May 1973.

[3] T. Briscoe, J. Carroll. Generalized Probabilistic LR Parsing of Natural Language (Corpora) with Unification-Based Grammars. In: *Computational Linguistics*, Vol. 19, No. 1.

[4] K. Lari and S.J. Young. Applications of Stochastic Context-free Grammars Using the Inside-Outside Algorithm. In:*Computer Speech and Language*,Vol. 5,pp. 237-257,1991.

[5] H.W.L. ter Doest. *Stochastic Grammars: Consistency and Inference.* M. Sc. Thesis, University of Twente, Enschede, in preparation, The Netherlands.

[6] T.E. Harris. *The Theory of Branching Processes.* Springer-Verlag (Berlin and New York), 1963.

[7] F. Jelinek, J.D. Lafferty. Computation of the Probability of Initial Substring Generation by Stochastic Context-Free Grammars. In: *Computational Linguistics*, Vol. 17, No. 3.

[8] A. Salomaa. Probabilistic and Weighted Grammars. In: *Information and Sciences*, Vol. 15 (1969), pp. 529-544.

[9] C.S. Wetherell. Probabilistic Languages: A Review and Some Open Questions. In: *Computing Surveys* , Vol. 12, No. 4, pp. 361-379, December 1980.