Two-level probabilistic grammars for natural language parsing
Infante Lopez, G.G.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Appendix B

Revising the STOP Symbol for Markov Rules

B.1 The Importance of the STOP Symbol

In this section we discuss Section 2.4.1 of Collins PhD thesis, where he discusses the importance of the STOP symbol for generating Markovian sequences. The idea is that any sequence of strings that is generated by a Markovian process should end in a STOP symbol. He argues that without the STOP symbol the probability distribution generated by a Markovian process over finite strings is not really a probability distribution.

We argue that the even though the STOP symbol is indeed important, the justification Collins provides is not fully correct. We show that the mere existence of the STOP symbol is not enough to guarantee consistency. We argue that probability distributions generated by Markovian process over finite sequences of symbols should be thought of as probability distributions defined over infinite sequences of symbols instead of probability distributions defined over finite sequences.

In Section B.2 we present Collins’s explanation on STOP symbols; in Section B.3 we provide background definitions on Markov chains; in Section B.4 we use Markov chains for rethinking the importance of STOP symbols, and in Section B.5 we conclude the appendix.

B.2 Collins’s Explanation

Suppose we want to assign a probability $p$ to sequences of symbols $w_1, w_2, w_3, \ldots, w_n$, where each symbol $w_i$ belongs to a finite alphabet $\Sigma$.

We first rewrite the probability of the given sequence using the chain rule of prob-
Appendix B. Revising the STOP Symbol for Markov Rules

Abilities as:

\[ P(w_1, w_2, w_3, \ldots, w_n) = \prod_{i=1}^{n} P(w_i|w_1, \ldots, w_{i-1}). \]  \hspace{1cm} (B.1)

Second, we use \( m \)-order Markovian independence assumptions and the equation above becomes:

\[ P(w_1, w_2, w_3, \ldots, w_n) = \prod_{i=1}^{n} P(w_i|w_1, \ldots, w_{i-m}). \]  \hspace{1cm} (B.2)

Collins (1999, page 46) argues that such a definition of probabilities is not correct: the problems arise because \( n \), the sentence length, is variable. In his argumentation, he states that Equation B.1 would be correct if the event space under consideration would have been the space of \( n \)-dimensional vectors \( \Sigma^n \) instead of the set of all strings in the language \( \Sigma^* \). Writing the probability under consideration as \( P(w_1, w_2, w_3, \ldots, w_n) \) implies that \( \Sigma^n \) is the event space. To avoid this confusion he writes the probability of a sequence \( \langle w_1, w_2, \ldots, w_n \rangle \) as \( P(\langle w_1, w_2, \ldots, w_n \rangle) \): the angled braces imply that \( \langle w_1, w_2, \ldots, w_n \rangle \) is a sequence of variable length rather than an \( n \)-dimensional vector.

He states that in the case of speech recognition, the length \( n \) of strings is often large and that STOP probabilities in that case may not be too significant. In Collins’s use of Markov processes, the sequences under consideration are typically of length 0, 1 or 2, and for this case the STOP probabilities certainly become important.

He presents the following example to support his point on the failing of Equation B.1

B.2.1. Example. Consider the following.

- Assume \( \Sigma = \{a, b\} \), and therefore that \( \Sigma^* \) is \( \{\epsilon, a, b, aa, bb, ab, bb, \ldots\} \).
- Assume that we will model the probability over \( \Sigma^* \) with a 0’th order Markov process, with parameters \( P(a) = P(b) = 0.5 \)

We can now calculate the probability of several strings using the formula in Equation B.2: \( P(\langle a \rangle) = 0.5, P(\langle b \rangle) = 0.5, P(\langle aa \rangle) = 0.5^2, P(\langle bb \rangle) = 0.25 \) and so on. We already see from these 4 probabilities that the sum, over the event space will be greater than 1: \( P(\langle a \rangle) + P(\langle b \rangle) + P(\langle aa \rangle) + P(\langle bb \rangle) = 1.5! \). An additional problem is that the probability of the empty string, \( P(\langle \rangle) \), where \( n = 0 \), is undefined.

Collins argues that adding STOP symbols fixes this inconsistency. He adds STOP symbols with the parameters of the Markov process modified to include. For example, let \( P(a) = P(b) = 0.25, P(\text{STOP}) = 0.5 \). In this case we have \( P(\langle \text{STOP} \rangle) = 0.5, P(\langle a\text{STOP} \rangle) = 0.25 \times 0.5 = 0.125, P(\langle b\text{STOP} \rangle) = 0.25 \times 0.5 = 0.125, P(\langle aa\text{STOP} \rangle) = 0.25^2 \times 0.5 = 0.03125, P(\langle bb\text{STOP} \rangle) = 0.03125 \) and so on. Thus far the sum of
probabilities does not exceed 1, and the distribution looks much better behaved. We can prove that the sum over all sequences is 1 by noting that the probability of any sequence of length \( n \) is \( 0.25^n \cdot 0.5 \), and that there are \( 2^n \) sequences of length \( n \), giving:

\[
\sum_{W \in \Sigma} \sum_{n=0}^{\infty} 2^n \cdot 0.25^n \cdot 0.5 = \sum_{n=0}^{\infty} 0.5^n \cdot 0.5 = \sum_{n=0}^{\infty} 0.5^{n+1} = \sum_{n=0}^{\infty} 0.5^n = 1.
\]

In a 0'th order Markov process the distribution over length of strings is related directly to \( P(\text{STOP}) \) — the probability of a string having length \( n \) is the probability of generating \( n \) non-STOP symbols followed by the STOP symbol:

\[
P(\text{length} = n) = (1 - P(\text{STOP}))^n \cdot p(\text{STOP}).
\]

With higher order Markov processes, where the probability is conditioned on previously generated symbols, the conditional probability \( P(\text{STOP}|w_{i-m}, \ldots, w_{i-1}) \) encodes the preference for certain symbols or sequences of symbols to end or not to end a sentence. For example, if we were building a bigram (1st order Markov) model of English we would expect the word the to end a sentence very rarely, and the corresponding parameter \( P(\text{STOP}|\text{the}) \) to be very low. The STOP symbol not only ensures the probability distributions to be well defined, but also to have useful interpretation.

## B.3 Background on Markov Chains

Markovian processes like the one described in Example B.2.1 are better described through Markov chains (Taylor and Karlin, 1998).

### B.3.1 Definition. A Markov chain is a 3-tuple \( M = (W, P, I) \) where \( W \) is a set of states, \( P \) is a real \( |W| \times |W| \) matrix with entries in \( \mathbb{R} \), such that \( \sum_{j=1}^{|W|} p_{ij} = 1 \); \( p_{ij} \) is the probability of jumping from state \( i \) to state \( j \), and \( I \) is \( |W| \)-dimensional vector defining the initial probability distribution.
B.3.2. **Example.** Let $M = (W, P, I)$ be a Markov chain where $W = \{a, b\}$.

$$
P = \begin{pmatrix}
a & b \\
0.5 & 0.5 \\
0.5 & 0.5
\end{pmatrix}
$$

and $I = (0.5, 0.5)^t$. The graphical representation for this Markov chain is in Figure B.1

![Graphical representation of the Markov chain](image)

**Figure B.1:** A graphical representation of the model in Example B.2.1

Markov chains may be viewed as discrete stochastic processes. A discrete stochastic process is a distribution over an infinite sequence of the random variables, each taking a value out of a finite set. We say that the process is *Markovian* if the outcome of a particular random variable in the sequence depends only on its two neighbors (the one before it and the one after it in the sequence).

The following definition formalizes this idea.

**B.3.3. Definition.** A stochastic process $\{W_0, W_1, \ldots, W_n, \ldots\}$ at consecutive points of observations $0, 1, \ldots, n, \ldots$ is a *discrete Markov process* if, for all $n \in \mathbb{N}$, $w_n \in W$

$$
P(W_{n+1} = w_{n+1} | W_n = w_n, W_{n-1} = w_{n-1}, \ldots, W_0 = w_0) = P(W_{n+1} = w_{n+1} | W_n = w_n)
$$

(B.3)  

(B.4)

Let $W = \{w_1, \ldots, w_n\}$. The quantities

$$
p_{ij} = P(W_{n+1} = w_j | W_n) = P(W_1 = w_j | W_0 = w_i)
$$

are known as the *transition probabilities* of the Markov process.

**B.3.4. Definition.** A state $W$ in a Markov chain is called *absorbing* if it does not have any outgoing arc, or, equivalently, every outgoing arc points to the state itself.
Markov chains have been widely used in software modeling for describing probability distributions over infinite sequences of states (Infante-Lopez et al., 2001). Under these approach, every string accepted by a Markov chain is in fact an infinite string, and it describes an infinite path in the Markov chain. To define the probability distribution, sets of infinite strings are used as a building block. Each block is characterized by a finite string α, and it contains all infinite strings starting with prefix α that are accepted by the Markov chain. The probability assigned to an α block, by a Markov chain $M$ is defined as the probability of traversing the path described by $\alpha$ in $M$. The probability associated to a path $X_1, \ldots, X_k$ is defined as $I(X_1) \times \prod_{i=1}^{k-1} P(X_i, X_{i+1})$.

The probability of a finite sequence $\alpha$ of length $n$ is defined as the probability of the infinite sequence $\alpha \beta$ where $\beta$ is an infinite sequence consisting of the STOP symbol, i.e., $\beta = \text{STOP}^\omega$. The intuition underlying this definition is that the Markov chains reach a state from which it can not leave, and consequently, the $\beta$ cycles for ever in the same state.

Under this perspective, the Markov chain in Example B.2.1 does not generate any distribution over finite sequences. But it does generate a distribution over infinite sequences. It is interesting to note that every single infinite path receives probability zero, but sets of infinite paths do receive probability values greater than zero. This situation is comparable to probability distributions over real numbers, where each single number receives a probability zero but where subsets of real numbers receive non-zero probability values.

### B.4 The STOP Symbol Revisited

Collins Example B.2.1 has a direct translation into a Markov chains. We can define a Markov chain with states $W = \{a, b\}$, and transition matrix

$$P(W_{n+1} = w_{n+1}|W_n = w_n, W_{n-1} = w_{n-1}, \ldots, W_0 = w_0) =$$

and initial distributions $I = (0.5, 0.5)$. Figure B.1 presents this Markov chain in a graphical way. This Markov chain generates a well defined probability distribution over infinite sequences of states, and since there is no absorbing state, it cannot be used for deriving probability distributions over finite sequences.

The solution suggested by Collins adds a new state named STOP. Again, the solution can be described using a zero order Markov chain; the corresponding chain is pictured in Figure B.2.

As Collins suggests, the probability of the STOP state is 0.5, since it is a zero order Markov model it has two fundamental properties: first, all incoming edges have the
Appendix B. Revising the STOP Symbol for Markov Rules

Figure B.2: A zero order Markov chain using the STOP symbol

same probability and second there is a directed arc connecting every pair of states. The resulting Markov chain has a serious problem. It assigns probability mass to strings like “aaSTOPabbSTOP”, i.e., strings containing the STOP symbol more than once.

The situation we have so far is that, even though Collins’s solution ruled out probability distributions over “correct strings” to sum up above one, the solution proposed produces distributions that sum up below one. We would like to disallow this kind of distributions, a way to solve this particular problem is to modify Figure B.2 into Figure B.3. This Markov chain is a first order Markov model, even more it can not be defined using 0 order Markov chains.

Figure B.3: A Markov chain with the STOP state absorbing.

Actually, it seems that Collins assumes that the STOP state is an absorbing state. His model is a zero order Markov chain plus the condition that generation stops once the STOP symbol has been generated. His solution can be restated as “there has to be a
STOP symbol and it has to be absorbing."

But again, this is not a proper solution, the mere presence an absorbing STOP symbol is not enough to rule out inconsistent models. There can be Markov chains with a STOP symbol, that still produce wrong distributions; an example is shown in Figure B.4.

Figure B.4: A Markov chain with an absorbing STOP symbol that defines an inconsistent probability distribution over finite strings.

The model assigns probability 0.5 to the infinity sequence $w_0w_1w_2w_2w_2w_2w_2\ldots$, and 0.5 to the finite string $w_0w_1w_3$. The problem is that the model will enter into the infinite cycle producing $w_4$ with probability 0.5.

Summing up: In principle the formalism presented by Collins has a potential probability inconsistency. But Collins’s model is still on the safe side because his Markov models are learned from data and the following lemma applies:

**B.4.1. Lemma.** *A Markov model learned from strings augmented with the STOP symbol generate consistent probability distributions over finite strings.*

The lemma has already been proven in Section 4.2.2.

**B.5 Conclusions**

We presented a different perspective for the presence of the STOP symbols in the generation of finite sequences of strings. We show that formally a Markov chain of order zero can not have an absorbing STOP state. We show that Collins explanation of STOP symbols is not fully correct and we show that independently of his explanation, the Markov chains he induces are consistent.