# A Structured Language Model

**Ciprian Chelba**
The Johns Hopkins University
CLSP, Barton Hall 320
3400 N. Charles Street, Baltimore, MD-21218
chelba@jhu.edu

## Abstract

The paper presents a language model that develops syntactic structure and uses it to extract meaningful information from the word history, thus enabling the use of long distance dependencies. The model assigns probability to every joint sequence of words–binary-parse-structure with headword annotation. The model, its probabilistic parametrization, and a set of experiments meant to evaluate its predictive power are presented.

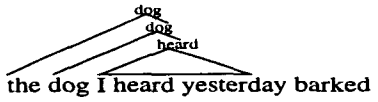


Figure 1: Partial parse

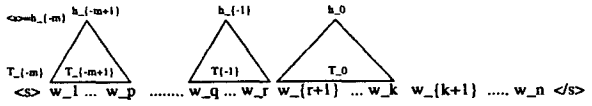


Figure 2: A word-parse k-prefix

## 1 Introduction

The main goal of the proposed project is to develop a language model(LM) that uses syntactic structure. The principles that guided this proposal were:

- the model will develop syntactic knowledge as a built-in feature; it will assign a probability to every joint sequence of words–binary-parse-structure;
- the model should operate in a left-to-right manner so that it would be possible to decode word lattices provided by an automatic speech recognizer.

The model consists of two modules: a next word *predictor* which makes use of syntactic structure as developed by a *parser*. The operations of these two modules are intertwined.

## 2 The Basic Idea and Terminology

Consider predicting the word barked in the sentence:

`the dog I heard yesterday barked again.`

A 3-gram approach would predict barked from (heard, yesterday) whereas it is clear that the predictor should use the word dog which is outside the reach of even 4-grams. Our assumption is that what enables us to make a good prediction of barked is the syntactic structure in the past. The correct *partial parse* of the word history when predicting barked is shown in Figure 1. The word dog is called the *headword* of the *constituent* ( the (dog (...))) and dog is an *exposed headword* when predicting barked — topmost headword in the largest constituent that contains it. The syntactic structure in the past filters out irrelevant words and points to the important ones, thus enabling the use of long distance information when predicting the next word. Our model will assign a probability $P(W,T)$ to every sentence $W$ with every possible binary branching parse $T$ and every possible headword annotation for every constituent of $T$. Let $W$ be a sentence of length $l$ words to which we have prepended <s> and appended </s> so that $w_0$ =<s> and $w_{l+1}$ =</s>. Let $W_k$ be the word k-prefix $w_0 \ldots w_k$ of the sentence and $W_k T_k$ the *word-parse k-prefix*. To stress this point, a word-parse k-prefix contains only those binary trees whose span is completely included in the word k-prefix, excluding $w_0$ =<s>. Single words can be regarded as root-only trees. Figure 2 shows a word-parse k-prefix; h_0 .. h_{-m} are the *exposed headwords*. A *complete parse* — Figure 3 — is any binary parse of the $w_1 \ldots w_l$ </s> sequence with the restriction that </s> is the only allowed headword.

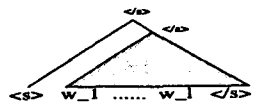


Figure 3: Complete parse

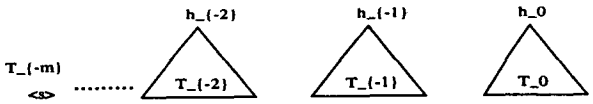


Figure 4: Before an adjoin operation

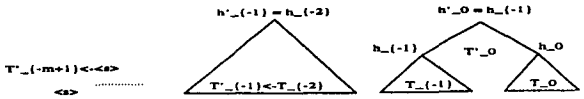


Figure 5: Result of adjoin-left

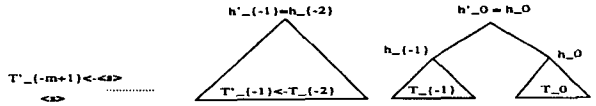


Figure 6: Result of adjoin-right

Note that $(w_1 \ldots w_l)$ *needn't* be a constituent, but for the parses where it is, there is no restriction on which of its words is the headword.

The model will operate by means of two modules:

- PREDICTOR predicts the next word $w_{k+1}$ given the word-parse k-prefix and then passes control to the PARSER;
- PARSER grows the already existing binary branching structure by repeatedly generating the transitions `adjoin-left` or `adjoin-right` until it passes control to the PREDICTOR by taking a `null` transition.

The operations performed by the PARSER ensure that all possible binary branching parses with all possible headword assignments for the $w_1 \ldots w_k$ word sequence can be generated. They are illustrated by Figures 4-6. The following algorithm describes how the model generates a word sequence with a complete parse (see Figures 3-6 for notation):

```
Transition t;          // a PARSER transition
generate <s>;
do{
  predict next_word;                  //PREDICTOR
  do{                                 //PARSER
    if(T_{-1} != <s> )
       if(h_0 == </s>)      t = adjoin-right;
       else t = {adjoin-{left,right}, null};
    else    t = null;
  }while(t != null)
}while(!(h_0 == </s> && T_{-1} == <s>))
t = adjoin-right; // adjoin <s>; DONE
```

It is easy to see that any given word sequence with a possible parse and headword annotation is generated by a unique sequence of model actions.

## 3 Probabilistic Model

The probability $P(W,T)$ can be broken into:
$P(W,T) = \prod_{k=1}^{l+1}[P(w_k/W_{k-1}T_{k-1}) \cdot \prod_{i=1}^{N_k} P(t_i^k/w_k, W_{k-1}T_{k-1}, t_1^k \ldots t_{i-1}^k)]$ where:

- $W_{k-1}T_{k-1}$ is the word-parse $(k-1)$-prefix
- $w_k$ is the word predicted by PREDICTOR
- $N_k - 1$ is the number of adjoin operations the PARSER executes before passing control to the PREDICTOR (the $N_k$-th operation at position k is the `null` transition); $N_k$ is a function of $T$
- $t_i^k$ denotes the i-th PARSER operation carried out at position k in the word string;

$t_i^k \in \{$`adjoin-left`,`adjoin-right`$\}, i < N_k$ ,
$t_i^k =$`null`$, i = N_k$
Our model is based on two probabilities:

$$P(w_k/W_{k-1}T_{k-1}) \quad (1)$$

$$P(t_i^k/w_k, W_{k-1}T_{k-1}, t_1^k \ldots t_{i-1}^k) \quad (2)$$

As can be seen $(w_k, W_{k-1}T_{k-1}, t_1^k \ldots t_{i-1}^k)$ is one of the $N_k$ word-parse k-prefixes of $W_kT_k, i = \overline{1, N_k}$ at position $k$ in the sentence.

To ensure a proper probabilistic model we have to make sure that (1) and (2) are well defined conditional probabilities and that the model halts with probability one. A few provisions need to be taken:

- $P($`null`$/W_kT_k) = 1$, if `T_{-1} == <s>` ensures that `<s>` is adjoined in the last step of the parsing process;
- $P($`adjoin-right`$/W_kT_k) = 1$, if `h_0 == </s>` ensures that the headword of a complete parse is `</s>`;
- $\exists \epsilon > 0 s.t.\ P(w_k=$`</s>`$/W_{k-1}T_{k-1}) \geq \epsilon, \forall W_{k-1}T_{k-1}$ ensures that the model halts with probability one.

### 3.1 The first model

The first term (1) can be reduced to an $n$-gram LM, $P(w_k/W_{k-1}T_{k-1}) = P(w_k/w_{k-1} \ldots w_{k-n+1})$.

A simple alternative to this degenerate approach would be to build a model which predicts the next word based on the preceding p-1 *exposed headwords* and n-1 words in the history, thus making the following equivalence classification:
$[W_kT_k] =$ {`h_0 .. h_{-p+2}`,$w_{k-1}..w_{k-n+1}$}.

The approach is similar to the trigger LM(Lau93), the difference being that in the present work triggers are identified using the syntactic structure.

### 3.2 The second model

Model (2) assigns probability to different binary parses of the word k-prefix by chaining the elementary operations described above. The workings of the PARSER are very similar to those of Spatter (Jelinek94). It can be brought to the full power of Spatter by changing the action of the `adjoin` operation so that it takes into account the terminal/nonterminal labels of the constituent proposed by `adjoin` and it also predicts the nonterminal label of the newly created constituent; PREDICTOR will now predict the next word along with its POS tag. The best equivalence classification of the $W_kT_k$ word-parse k-prefix is yet to be determined. The Collins parser (Collins96) shows that dependency-grammar–like bigram constraints may be the most adequate, so the equivalence classification $[W_kT_k]$ should contain at least `{h_0, h_{-1}}`.

## 4 Preliminary Experiments

Assuming that the correct partial parse is a function of the word prefix, it makes sense to compare the word level perplexity(PP) of a standard n-gram LM with that of the $P(w_k/W_{k-1}T_{k-1})$ model. We developed and evaluated four LMs:

- 2 bigram LMs $P(w_k/W_{k-1}T_{k-1}) = P(w_k/w_{k-1})$ referred to as W and w, respectively; $w_{k-1}$ is the previous (word, POStag) pair;
- 2 $P(w_k/W_{k-1}T_{k-1}) = P(w_k/h_0)$ models, referred to as H and h, respectively; $h_0$ is the previous exposed (headword, POS/non-term tag) pair; the parses used in this model were those assigned manually in the Penn Treebank (Marcus95) after undergoing headword percolation and binarization.

All four LMs predict a word $w_k$ and they were implemented using the Maximum Entropy Modeling Toolkit[1] (Ristad97). The constraint templates in the {W,H} models were:

```
4 <= <*>_<*> <?>; 2 <= <?>_<*> <?>;
2 <= <?>_<?> <?>; 8 <= <*>_<?> <?>;
```

and in the {w,h} models they were:

```
4 <= <*>_<*> <?>; 2 <= <?>_<*> <?>;
```

<*> denotes a *don't care* position, <?>_<?> a (word, tag) pair; for example, `4 <= <?>_<*> <?>` will trigger on all ((word, *any tag*), predicted-word) pairs that occur more than 3 times in the training data. The sentence boundary is not included in the PP calculation. Table 1 shows the PP results along with the number of parameters for each of the 4 models described .

| LM | PP | param | LM | PP | param |
|---|---|---|---|---|---|
| W | 376 | 208487 | w | 419 | 103732 |
| H | 312 | 206540 | h | 410 | 102437 |

Table 1: Perplexity results

[1] ftp://ftp.cs.princeton.edu/pub/packages/memt

## 5 Acknowledgements

The author thanks to Frederick Jelinek, Sanjeev Khudanpur, Eric Ristad and all the other members of the Dependency Modeling Group (Stolcke97), WS96 DoD Workshop at the Johns Hopkins University.

## References

Michael John Collins. 1996. A new statistical parser based on bigram lexical dependencies. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, 184-191, Santa Cruz, CA.

Frederick Jelinek. 1997. Information extraction from speech and text — course notes. The Johns Hopkins University, Baltimore, MD.

Frederick Jelinek, John Lafferty, David M. Magerman, Robert Mercer, Adwait Ratnaparkhi, Salim Roukos. 1994. Decision Tree Parsing using a Hidden Derivational Model. In *Proceedings of the Human Language Technology Workshop*, 272-277. ARPA.

Raymond Lau, Ronald Rosenfeld, and Salim Roukos. 1993. Trigger-based language models: a maximum entropy approach. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, volume 2, 45-48, Minneapolis.

Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz. 1995. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313-330.

Eric Sven Ristad. 1997. Maximum entropy modeling toolkit. Technical report, Department of Computer Science, Princeton University, Princeton, NJ, January 1997, v. 1.4 Beta.

Andreas Stolcke, Ciprian Chelba, David Engle, Frederick Jelinek, Victor Jimenez, Sanjeev Khudanpur, Lidia Mangu, Harry Printz, Eric Sven Ristad, Roni Rosenfeld, Dekai Wu. 1997. Structure and Performance of a Dependency Language Model. In *Proceedings of Eurospeech'97*, Rhodes, Greece. To appear.