Lexical Heads, Phrase Structure and the Induction of Grammar

Carl de Marcken
MIT Artificial Intelligence Laboratory
NE43-804
545 Technology Square
Cambridge, MA, 02139, USA
cgdemarc@ai.mit.edu

Summary

Acquiring linguistically plausible phrase-structure grammars from ordinary text has proven difficult for standard induction techniques, and researchers have turned to supervised training from bracketed corpora. We examine why previous approaches have failed to acquire desired grammars, concentrating our analysis on the inside-outside algorithm (Baker, 1979), and propose that with a representation of phrase structure centered on head relations such supervision may not be necessary.

1. INTRODUCTION

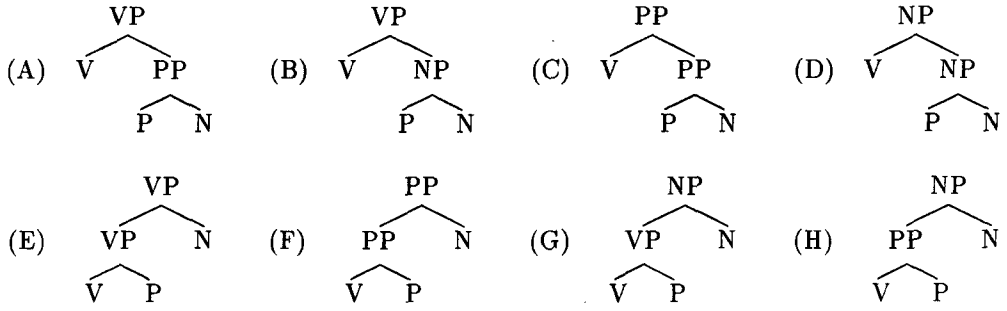
Researchers investigating the acquisition of phrase-structure grammars from raw text have had only mixed success. In particular, unsupervised learning techniques, such as the inside-outside algorithm (Baker, 1979) for estimating the parameters of stochastic context-free grammars (SCFGs), tend to produce grammars that structure text in ways contrary to our linguistic intuitions. One effective way around this problem is to use hand-structured text like the Penn Treebank (Marcus, 1991) to train the learner: (Pereira and Schabes, 1992) demonstrate that the inside-outside algorithm can learn grammars effectively given such constraint; from a bracketed corpus (Brill, 1993) successfully learns rules that iteratively transform a default phrase-structure into a better one for a particular sentence.

The necessity of bracketed corpora for training is grating to our sensibilities, for several reasons. First, bracketed corpora are not easy to come by. Second, there is a sense that in learning from them, little of interest is going on. In the case of the acquisition of stochastic context-free grammars, the parameters can be read off of a fully-bracketed corpus by simply counting. Finally, the inability of current models to learn (without supervision) the parameters we desire suggests that our models are mismatched to the problem.

This paper examines why some previous approaches have failed to acquire desired grammars without supervision, and proposes that with a different conception of phrase-structure supervision might not be necessary. In particular, we examine some reasons why SCFGs are poor models to use for learning human language, especially when combined with the inside-outside algorithm. We argue that head-driven grammatical formalisms like dependency grammars (Melčuk, 1988) or link grammars (Sleator and Temperley, 1991) are better suited to the task.

2. LINGUISTIC AND STATISTICAL BASIS OF PHRASE STRUCTURE

Let us look at a particular example. In English, the word sequence *"walking on ice"* is generally assumed to have an internal structure similar to (A).[1]



Why (A) and not one of (B-H)? An introductory linguistics book might suggest the following answers:

- *on ice* can move and delete as one unit, whereas *walking on* can not. Thus, *"it is on ice that I walked"* and *"it is walking that I did on ice"* and *"it is ice that I walked on"* are sentences but there in no equivalent form for relocating *walking on*. Similarly, *"they walked and jumped on ice"* is grammatical but *"they walked on and jumped on ice"* is awkward. Therefore, if movement and conjunction is of single constituents, phrase-structures (A-D) explain this evidence but (E-H) do not.
- In languages like German where case is overtly manifested in affix and determiner choice, the noun *ice* clearly receives case from the preposition rather than the verb. It seems to make for a simpler theory of language if case is assigned through the government relation, which holds between the preposition and noun in (A-D) but not in (E-H).
- The phrase *walking on ice* acts like a verb: it can conjoin with a verb (*"John walked on ice and sang"*), and takes verbal modifiers (*"John walked on ice slowly"*). So it makes little sense to call it a prepositional phrase or noun phrase, as in (C) or (D). *on ice* does not behave as a noun, so (A) is a better description than (B).

These deductive steps leading to (A) require some assumptions about language: that constituent structure and category labels introduce specific constraints on sentence building operations, and that the range of hypothetical grammars is small (our enumeration A-H was over grammars of binary rules where the category of a phrase is tied to the category of one of its constituents, its head).

[1] We will be deliberately vague about what such dominance and precedence relations represent; obviously different researchers have very different conceptions about the relevence and implications of heirarchical phrase-structure. The specific use of the representations is somewhat irrelevent to our immediate discussion, though various interpretaions will be discussed throughout the paper.

Statistical phrase-structure models of language[2], such as SCFGs, are motivated by different assumptions about language, principally that a phrase grouping several words is a constraint on co-occurrence that makes it possible to better predict one of those words given another. In terms of language acquisition and parsing, if we assume that a sequence of words has been generated from a phrase-structure grammar, it suggests that we can recover internal structure by grouping subsequences of words with high mutual information. This is the approach taken by (Magerman and Marcus, 1990) for parsing sentences, who use mutual information rather than a grammar to reconstruct phrase-structure. The hope is that by searching for a phrase-structure or phrase-structure grammar that maximizes the likelihood of an observed sequence, we will find the generating structure or grammar itself.

Unfortunately, there is anecdotal and quantitative evidence that simple techniques for estimating phrase-structure grammars by minimizing entropy do not lead to the desired grammars (grammars that agree with structure (A), for instance). (Pereira and Schabes, 1992) explore this topic, demonstrating that a stochastic context free grammar trained on part-of-speech sequences from English text can have an entropy as low or lower than another but bracket the text much more poorly (tested on hand-annotations). And (Magerman and Marcus, 1990) provide evidence that grouping sub-sequences of events with high mutual information is not always a good heuristic; they must include in their parsing algorithm a list of event sequences (such as noun-preposition) that should not be grouped together in a single phrase, in order to prevent their method from mis-bracketing. To understand why, we can look at an example from a slightly different domain.

(Olivier, 1968) seeks to acquire a lexicon from unsegmented (spaceless) character sequences by treating each word as a stochastic context free rule mapping a common nonterminal (call it W) to a sequence of letters; a sentence is a sequence of any number of words and the probability of a sentence is the product over each word of the probability of W expanding to that word. Learning a lexicon consists of finding a grammar that reduces the entropy of a training character sequence. Olivier's learning algorithm soon creates rules such as W $\Rightarrow$ THE and W $\Rightarrow$ TOBE. But it also hypothesizes words like *edby*. *edby* is a common English character sequence that occurs in passive constructions like *"She was passed by the runner"*. Here *-ed* and *by* occur together not because they are part of a common word, but because English syntax and semantics places these two morphemes side-by-side. At a syntactic level, this is exactly why the algorithm of (Magerman and Marcus, 1990) has problems: English places prepositions after nouns not because they are in the same phrase, but because prepositional phrases often adjoin to noun phrases. Any greedy algorithm (such as (Magerman and Marcus, 1990) and the context-free grammar induction method of (Stolcke, 1994)) that builds phrases by grouping events with high mutual information will consequently fail to derive linguistically-plausible phrase structure in many situations.

3. INCORPORATING HEADEDNESS INTO LANGUAGE MODELS

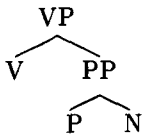
The conclusion of the above section might lead us to is that basing phrase-structure grammar induction on minimization of entropy is a poor idea. However, in this paper we will not discuss whether statistical induction is the proper way to view language acquisition: our current goal is only to better understand why current statistical methods produce the "wrong" answer and to

[2]While this paper concentrates on the acquisition of syntax, similar or identical statistical models to those discussed here have been used to acquiring words and morphemes from sequences of characters (Olivier, 1968; Wolff, 1982; Brent, 1993; Cartwright and Brent, 1994) and syllables from phonemes (Ellison, 1992), among other language applications.

explore ways of fixing them.

Let us look again at (A), reproduced below, and center discussion on an extended stochastic context-free grammar model in which a binary context-free rule Z $\Rightarrow$ A B with terminal parts-of-speech on the right hand side first generates a word a chosen from a distribution $p_{\mathrm{A}}(a)$, then generates a word b from a distribution $p_{\mathrm{B}}^{a}(b)$.[3] If we call these two random variables A and B, then the entropy of the sequence AB is $H(A) + H(B|A) = H(A) + H(B) - I(A,B)$ (where $H(X)$ is the entropy of a random variable X and $I(X,Y)$ is the mutual information between two random variables X and Y). The point here is that using such a context free rule to model a sequence of two words reduces the entropy of the language from a model that treats the two words as independent, by precisely the mutual information between the two words.

(A)



In English, verbs and prepositions in configuration (A) are closely coupled semantically, probably more closely than prepositions and nouns, and we would expect that the mutual information between the verb and preposition would be greater than between the preposition and noun, and greater still than between the verb and the noun.

$$I(V,P) > I(P,N) > I(V,N)$$

Under our hypothesized model, structure (A) has entropy $H(V) + H(P) + H(N|P) = H(V) + H(P) + H(N) - I(P,N)$, which is higher than the entropy of structures (E-H), $H(V) + H(P) + H(N) - I(V,P)$, and we wouldn't expect a learning mechanism based on such a model to settle on (A).

However, this simple class of models only uses phrases to capture relations between adjacent words. In (A), it completely ignores the relation between the verb and the prepositional phrase, save to predict that a prepositional phrase (*any* prepositional phrase) will follow the verb. We modify our language model, assuming that nonterminals exhibit the distributional properties of their heads. We will write a phrase Z that is headed by a word z as $\langle \mathrm{Z}, z \rangle$. Each grammar rule will look like either $\langle \mathrm{Z'}, z \rangle \Rightarrow \langle \mathrm{Z}, z \rangle \langle \mathrm{Y}, y \rangle$ or $\langle \mathrm{Z'}, z \rangle \Rightarrow \langle \mathrm{Y}, y \rangle \langle \mathrm{Z}, z \rangle$ (abbreviated Z' $\Rightarrow$ Z Y and Z' $\Rightarrow$ Y Z) and the probability model is

$$\begin{aligned} p(\langle \mathrm{Z}, z \rangle \langle \mathrm{Y}, y \rangle | \langle \mathrm{Z'}, z' \rangle, \mathrm{Z'} \Rightarrow \mathrm{Z\ Y}) &= p_{\mathrm{Z}}(z|z') \cdot p_{\mathrm{Y}}^{z}(y|z') \cdot \delta(z,z') \\ &= p_{\mathrm{Y}}^{z}(y) \cdot \delta(z,z'). \end{aligned} \tag{1}$$

[3] Our notation here is that $p_{\mathrm{A}}(a)$ is the probability of word a being generated by a terminal part-of-speech A, and $p_{\mathrm{B}}^{a}(b)$ is the probability of the terminal part-of-speech B generating the word b given that previous word generated in the same phrase is a.

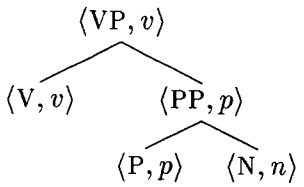
$$
\begin{aligned}
p(\langle \mathrm{Y}, y\rangle\langle \mathrm{Z}, z\rangle | \langle \mathrm{Z'}, z'\rangle, \mathrm{Z'} \Rightarrow \mathrm{Y\ Z}) &= p_{\mathrm{Y}}(y|z') \cdot p_{\mathrm{Z}}^{y}(z|z') \cdot \delta(z, z') \\
&= p_{\mathrm{Y}}(y|z') \cdot \delta(z, z') \\
&= p_{\mathrm{Y}}^{z}(y) \cdot \delta(z, z'). \qquad (2)
\end{aligned}
$$

Of course, this class of models is strongly equivalent to ordinary context free grammars. We could substitute, for every rule Z' $\Rightarrow$ Z Y, a large number of word-specific rules

$$\langle \mathrm{Z'}, z_i\rangle \Rightarrow \langle \mathrm{Z}, z_i\rangle\langle \mathrm{Y}, y_j\rangle$$

with probabilities $p(\mathrm{Z'} \Rightarrow \mathrm{Z\ Y}) \cdot p_{\mathrm{Y}}^{z_i}(y_j)$.

Using our new formalism, the head properties of (A) look like



and the entropy is

$$H(V) + H(P|V) + H(N|P) = H(V) + H(P) + H(N) - I(V, P) - I(P, N).$$

The grammar derived from (A) is optimal under this model of language, though (C), (F), and (H) are equally good. They could be distinguished from (A) in longer sentences because they pass different head information out of the phrase. In fact, the grammar model derived from (A) is as good as any possible model that does not condition N on V. Under this class of models there is no benefit to grouping two words with high mutual information together in the same minimal phrase; it is sufficient for both to be the heads of phrases that are adjacent at some level.

There is of course no reason why the best head-driven statistical model of a given language *must* coincide with a grammar derived by a linguist. The above class of models makes no mention of deletion or movement of phrases, and only information about the head of a phrase is being passed beyond that phrase's borders. The government-binding framework usually supposes that an inflection phrase is formed of inflection and the verb phrase. But the verb is likely to have a higher mutual information with the subject than inflection does. So it seems unlikely that this structure would be learned using our scheme. The effectiveness of the class of models can only be verified by empirical tests.

4. SOME EXPERIMENTS

We have built a stochastic, feature-based Earley parser (de Marcken, 1995) that can be trained using the inside-outside algorithm. Here we describe some tests that explore the interaction of the head-driven language models described above with this parser and training method.

For all the tests described here, we learn a grammar by starting with an exhaustive set of stochastic context-free rules of a certain form, and estimate probabilities for these rules from a test corpus. This is the same general procedure as used by (Lari and Young, 1990; Briscoe and Waegner, 1992; Pereira and Schabes, 1992) and others. For parts-of-speech Y and Z, the rules we include in our base grammar are

$$\begin{array}{lll} S \Rightarrow ZP & ZP \Rightarrow ZP\ YP & ZP \Rightarrow YP\ ZP \\ ZP \Rightarrow Z\ YP & ZP \Rightarrow YP\ Z & ZP \Rightarrow Z \end{array}$$

where S is the root nonterminal. As is ususal with stochastic context-free grammars, every rule has an associated probability, and the probabilities of all the rules that expand a single nonteminal must sum to one. Furthermore, each word and phrase has an associated head word (represented as a feature value that is propagated from the Z or ZP on the right hand side of the above rules to the left hand side). The parser is given the part of speech of each word.

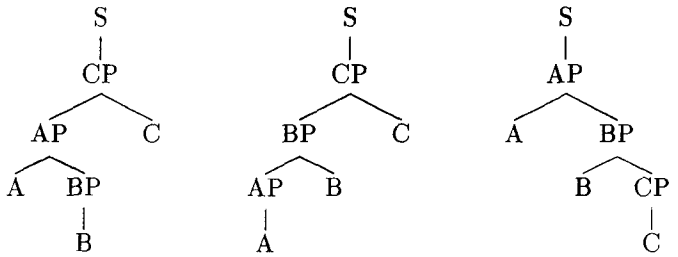
For binary rules, as per equations (1) and (2), the distribution of the non-head word is conditioned on the head (a bigram). Initially, all word bigrams are initialized to uniform distributions, and context-free rule probabilities are initialized to a small random perturbation of a uniform distribution.

4.1. A Very Simple Sentence

We created a test corpus of 1000 sentences, each 3 words long with a constant part-of-speech pattern ABC. Using 8 equally probable words per part-of-speech, we chose a word distribution over the sentences with the following characteristics:

$$I(A,B) = 1 \text{ bit.} \quad I(B,C) = 0.188 \text{ bits.} \quad I(A,C) = 0 \text{ bits.}$$

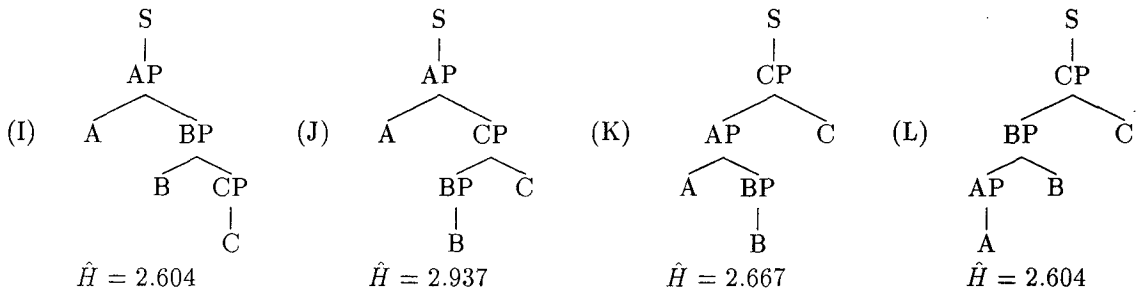
In other words, given knowledge of the first word in the sentence, predicting the second word is as difficult as guessing between four equally-likely words, and knowing the second word makes predicting the third as difficult as guessing between seven words. Knowing the first gives no information about the third. This is qualitatively similar to the distribution we assumed for verbs, nouns, and prepositions in configuration (A), and has entropy rate $\frac{3+(3-1)+(3-.188)}{3} = 2.604$ bits per word. Across 20 runs, the training algorithm converged to three different grammars:[4]



Count:	12	2	6
Cross-Entropy Rate:	2.648 bits/word	2.570 bits/word	2.570 bits/word

[4] *I.e.*, after the cross-entropy had ceased to decrease on a given run, the parser settled on one of these structures as the Viterbi parse of each sentences in the corpus. The cross-entropy rate of the two best grammars is lower than the source entropy rate because the corpus is finite and randomly generated, and has been be overfitted.

One fact is immediately striking: even with such simple sentences and rule sets, more often than not the inside-outside algorithm converges to a suboptimal grammar. To understand why, let us ignore recursive rules (ZP ⇒ ZP YP) for the moment. Then there are four possible parses of ABC (cross-entropy rate with source given below- lower is better model):



During the first pass of the inside-outside algorithm, assuming near-uniform initial rule probabilities, each of these parses will have equal posterior probabilities. They are equally probable because they use the same number of expansions[5] and because word bigrams are uniform at the start of the parsing process. Thus, the estimated probability of a rule after the first pass is directly proportional to how many of these parse trees the rule features in. The rules that occur more than one time are:

AP ⇒ A BP	(parses I,K)
CP ⇒ BP C	(parses J,L)
BP ⇒ B	(parses J,K)

Therefore, on the second iteration, these three rules will have higher probabilities than the others and will cause parses J and K to be favored over I and L (with K favored over J because $I(A,B) + I(A,C) > I(B,C) + I(A,C)$). It is to be expected then, that the inside-outside algorithm favors the suboptimal parse K: at its start the inside-outside algorithm is guided by tree counting arguments, not mutual information between words. This suggests that the inside-outside algorithm is likely to be highly sensitive to the form of grammar and how many different analyses it permits of a sentence.

Why, later, does the algorithm not move towards a global optimum? The answer is that the inside-outside algorithm is supremely unsuited to learning with this representation. To understand this, notice that to move from the initially favored parse (K) to one of the optimal ones (I and L), three nonterminals must have their most probable rules switched:

(K)	⟶	(L)
AP ⇒ A BP	⟶	AP ⇒ A
BP ⇒ B	⟶	BP ⇒ AP B
CP ⇒ AP C	⟶	CP ⇒ BP C

[5]This is why we can safely ignore recursive rules in this discussion. Any parse that involves one will have a bigger tree and be significantly less probable.

To simplify the present analysis, let us assume the probability of S $\Rightarrow$ CP is held constant at 1, and that the rules not listed above have probability 0. In this case, we can write the probabilities of the left three rule as p^{A}, p^{B} and p^{C} and the probabilities of the right rhree rules as $1-p^{\mathrm{A}}$, $1-p^{\mathrm{B}}$ and $1-p^{\mathrm{C}}$. Now, for a given sentence *abc* there are only two parses with non-zero probabilities, K and L. The probability of *abc* under parse K is $p^{\mathrm{A}}p^{\mathrm{B}}p^{\mathrm{C}}p(c)p(a|c)p(b|a)$, and the probability under parse L is $(1-p^{\mathrm{A}})(1-p^{\mathrm{B}})(1-p^{\mathrm{C}})p(c)p(b|c)p(a|b)$. Thus, the posterior probability of parse K is[6]

$$
\begin{aligned}
p(\mathrm{K}|abc) &= \frac{p^{\mathrm{A}}p^{\mathrm{B}}p^{\mathrm{C}}p(c)p(a|c)p(b|a)}{p^{\mathrm{A}}p^{\mathrm{B}}p^{\mathrm{C}}p(c)p(a|c)p(b|a)+(1-p^{\mathrm{A}})(1-p^{\mathrm{B}})(1-p^{\mathrm{C}})p(c)p(b|c)p(a|b)} \\
&= \frac{1}{1+\frac{(1-p^{\mathrm{A}})(1-p^{\mathrm{B}})(1-p^{\mathrm{C}})p(b|c)p(a|b)}{p^{\mathrm{A}}p^{\mathrm{B}}p^{\mathrm{C}}p(a|c)p(b|a)}} \\
&= \frac{1}{1+\frac{(1-p^{\mathrm{A}})(1-p^{\mathrm{B}})(1-p^{\mathrm{C}})p(c|b)}{p^{\mathrm{A}}p^{\mathrm{B}}p^{\mathrm{C}}p(c|a)}}
\end{aligned}
$$

Since the inside-outside algorithm reestimates p^{A}, p^{B} and p^{C} directly from the sums of the posterior probabilities of K and L over the corpus, the probability update rule from one iteration to the next is

$$
p^{\mathrm{A}}, p^{\mathrm{B}}, p^{\mathrm{C}} \leftarrow \frac{1}{1+\frac{(1-p^{\mathrm{A}})(1-p^{\mathrm{B}})(1-p^{\mathrm{C}})}{p^{\mathrm{A}}p^{\mathrm{B}}p^{\mathrm{C}}}\alpha}.
$$

where α is the mean value of $p(c|b)/p(c|a)$, $\frac{8}{7}$ in the above test. Figure 4.1 graphically depicts the evolution of this dynamical system. What is striking in this figure is that the inside-outside algorithm is so attracted to grammars whose terminals concentrate probability on small numbers of rules that it is incapable of performing real search. Instead, it zeros in on the nearest such grammar, only biased slightly by its relative merits. We now have an explanation for why the inside-outside algorithm converges to the suboptimal parse K so often: the first ignorant iteration of the algorithm biases the parameters towards K, and subsequently there is an overwhelming tendency to move to the nearest deterministic grammar. This is a strong indication that the algorithm is a poor choice for estimating grammars that have competing rule hypotheses.

4.2. Multiple Expansions of a Nonterminal

For this test, the sentences were four words long (ABCD), and we chose a word distribution with the following characteristics:

$I(A,B) = 1$ bit. $I(A,D) = 1$ bit. $I(C,D) = 0$ bits.
$I(A,C) = 1$ bit. $I(B,C) = 0$ bits. $I(B,D) = 0$ bits.

It might seem that a minimal-entropy grammar for this corpus would be

[6] In the following derivation, understand that for word bigrams $p(a|b) = p(b|a)$ because $p(a) = p(b) = \frac{1}{8}$.

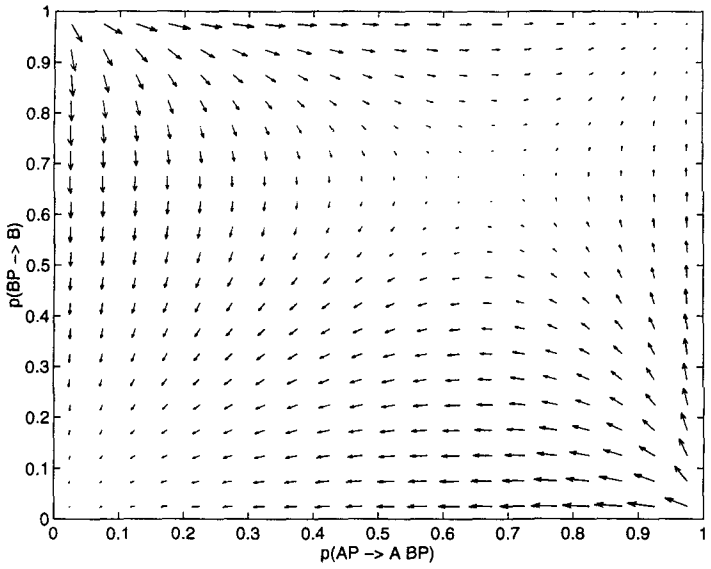


Figure 4.1: The dynamical properties of the inside-outside algorithm. The x-axis is p^{A} and the y-axis is p^{B}. The vectors represent the motion of the parameters from one iteration to the next when $\alpha = \frac{p(c|b)}{p(c|a)} = 2$ and $p^{\mathrm{C}} = .5$. Notice that the upper right corner (grammar K) and the lower left (grammar L) are stationary points (local maxima), and that the region of attraction for the global optimum L is bigger than for K, but that there is still a very substantial set of starting points from which the algorithm will converge to the suboptimal grammar. $\alpha = 2$ is plotted instead of $\alpha = \frac{8}{7}$ because this better depicts the asymmetry mutual information between words introduces; with $\alpha = \frac{8}{7}$ the two regions of attraction would be of almost equal area.

$$\begin{array}{lll} \mathrm{S} \Rightarrow \mathrm{DP} & \mathrm{DP} \Rightarrow \mathrm{AP\ D} & \mathrm{AP} \Rightarrow \mathrm{AP\ CP} \\ \mathrm{AP} \Rightarrow \mathrm{A\ BP} & \mathrm{CP} \Rightarrow \mathrm{C} & \mathrm{BP} \Rightarrow \mathrm{B} \end{array}$$

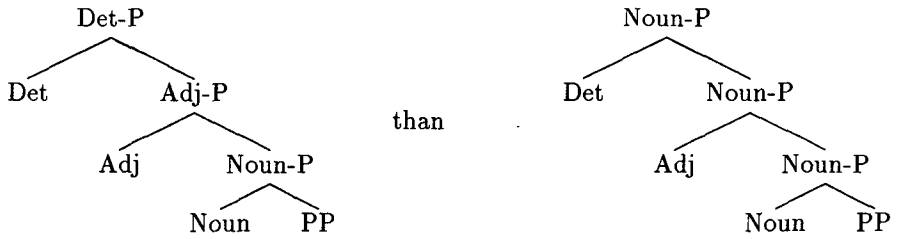
since this grammar makes the head A available to predict B, C, and D. Without multiple expansions rules for AP, it is impossible to get this. But the gain of one bit in word prediction is offset by a loss of at least two bits from uncertainty in the expansion of AP. Even if $p(\mathrm{AP} \Rightarrow \mathrm{A\ BP}) = p(\mathrm{AP} \Rightarrow \mathrm{AP\ CP}) = 1/2$, the probability of the structure ABCD under the above grammar is one-quarter that assigned by a grammar with no expansion ambiguity. So, the grammar

$$\begin{array}{lll} \mathrm{S} \Rightarrow \mathrm{DP} & \mathrm{DP} \Rightarrow \mathrm{CP\ D} & \mathrm{CP} \Rightarrow \mathrm{AP\ C} \\ \mathrm{AP} \Rightarrow \mathrm{A\ BP} & \mathrm{BP} \Rightarrow \mathrm{B} & \end{array}$$

assigns higher probabilities to the corpus, even though it fails to model the dependency between A and D. This is a general problem with SCFGs: there is no way to optimally model multiple ordered adjunction without increasing the number of nonterminals. Not surprisingly, the learning algorithm never converges to the recursive grammar during test runs on this corpus.

What broader implication does this deficiency of SCFGs have for context-free grammar based

language acquisition? It suggests if we were to estimate a grammar from English text, that the sequence Det Adj Noun PP is far more likely to get the interpretation



and therefore that, for many subject and object noun phrases, the noun will never enter into a bigram relationship with the verb. Obviously sufficient mutual information between nouns and verbs, adjectives, and determiners would force the global optimum to include multiple expansions of the Noun-P category, but it seems likely (given the characteristics of the inside-outside algorithm) that before such mutual information could be inferred from text, the inside-outside algorithm would enter a local optimum that does not pass the noun feature out.

4.3. Testing on the Penn Treebank

To test whether head-driven language models do indeed converge to linguistically-motivated grammars better than SCFGs, we replicated the experiment of (Pereira and Schabes, 1992) on the ATIS section of the Penn Treebank. The 48 parts-of-speech in the Treebank were collapsed to 25, resulting in 2550 grammar rules.

Word head features were created by assigning numbers a common feature; other words found in any case variation in the CELEX English-language database were given a feature particular to their lemma (thus mapping *car* and *cars* to the same feature); and all other (case-sensitve) words received their own unique feature. Treebank part-of-speech specifications were not used to constrain parses.

Bigrams were estimated using a backoff to a unigram (see (de Marcken, 1995)), and unigrams backing off to a uniform distribution over all the words in the ATIS corpus. The backoff parameter was not optimized. Sentences 25 words or longer were skipped.

We ran four experiments, training a grammar with and without bracketing and with and without use of features. Without features, we are essentially replicating the two experiments run by (Pereira and Schabes, 1992), except that they use a different set of initial rules (all 4095 CNF grammar rules over 15 nonterminals and the 48 Treebank terminal categories). Every tenth sentence of the 1129 sentences in the ATIS portion of the Treebank was set aside for testing. Training was over 1060 sentences (1017 of which 57 were skipped because of length), 5895 words, testing over 98 sentences (112, 14 skipped), 911 words.

After training, all but the 500 most probable rules were removed from the grammar, and probabilities renormalized. The statistics for these smaller grammars are given below.

Training	Grammar	Corpus	Perplexity	Bracketing
Bracketed	No Features	Train	55.68	90.1%
Bracketed	No Features	Test	95.15	88.5%
Unbracketed	No Features	Train	56.34	72.4%
Unbracketed	No Features	Test	92.91	72.7%
Bracketed	Features	Train	19.95	92.0%
Bracketed	Features	Test	68.88	90.7%
Unbracketed	Features	Train	19.31	73.3%
Unbracketed	Features	Test	72.12	74.8%

There are several notable qualities to these numbers. The first is that, in contrast to the results of (Pereira and Schabes, 1992), unbracketed training does improve bracketing performance (from a baseline of about 50% to 72.7% without features and 74.8% with features). Unfortunately, this performance is achieved by settling on an uninteresting right-branching rule set (save for sentence-final punctuation). Note that our figures for bracketed training match very closely to the 90.36% bracketing accuracy reported in their paper.

Of greater interest is that although use of head features improves bracketing performance, it does so only by an insignificant amount (though obviously it greatly reduces perplexity). There are many possible explanations for this result, but the two we prefer are that either the inside-outside algorithm, as might be expected given our arguments, failed to find a grammar that propagated head features optimally, or that there was insufficient mutual information in the small corpus for our enhancement to traditional SCFGs to have much impact.

We have replicated the above experiments on the first 2000 sentences of the Wall Street Journal section of the Treebank, which has a substantially different character than the ATIS text. However, the vocabulary is so much larger that is is not possible to gather useful statistics over such a small sample. The reason we have not tested extensively on much larger corpora is that, using head features but no bracketing constraint, statistics must be recorded for every word pair in every sentence. The number of such statistics grows quadratically with sentence length, and is prohibitive over large corpora using our current techniques. More recent experiments, however, indicate that expanding the corpus size by an order of magnitude has little affect on our results.

5. CONCLUSIONS

We have argued that there is little reason to believe SCFGs of the sort commonly used for grammar induction will ever converge to linguistically plausible grammars, and we have suggested a modification (namely, incorporating mutual information between phrase heads) that should help fix the problem. We have also argued that the standard context-free grammar estimation procedure, the inside-outside algorithm, is essentially incapable of finding an optimal grammar without bracketing help.

We now suggest that a representation that explicitly represents relations between phrase heads, such as link grammars (Sleator and Temperley, 1991), is far more amenable to language acquisition problems. Let us look one final time at the sequence V P N. There are only three words here, and therefore three heads. Assuming a head-driven bigram model as before, there are only three possibile anlayses of this sequence, which we write by listing the pairs of words that enter into predictive relationships:

Head Relations	Equivalent Phrase Structures
V-P, V-N	E,G
V-P, P-N	A,C,F,H
V-N, P-N	B,D

To map back into traditional phrase structure grammars, linking two heads X-Y is the same as specifying that there is some phrase XP headed by X which is a sibling to some phrase YP headed by Y. Of course, using this representation all of the optimal phrase structure grammars (A,C,F and H) are identical. Thus we have a representation which has factored out many details of phrase structure that are unimportant as far as minimizing entropy is concerned.

Simplifying the search space reaps additional benefits. A greedy approach to grammar acquisition that iteratively hypothesizes relations between the words with highest mutual information will first link V to P, then P to N, producing exactly the desired result for this example. And the distance in parse or grammar space between competing proposals is at most one relation (switching V-P to V-N, for instance), whereas three different rule probabilities may need to be changed in the SCFG representation. This suggests that learning algorithms based on this representation are far less likely to encounter local maximums. Finally, since what would have been multiple parse hypotheses are now one, a Viterbi learning scheme is more likely to estimate accurate counts. This is important, given the computational complexity of estimating long-distance word-pair probabilities from unbracketed corpora.

We have implemented a statistical parser and training mechanism based on the above notions, but results are too preliminary to include here. Stochastic link-grammar based models have been discussed (Lafferty et al., 1992) but the only test results we have seen (Della-Pietra et al., 1994) assume a very restricted subset of the model and do not explore the "phrase structures" that result from training on English text.

REFERENCES

James K. Baker. 1979. Trainable grammars for speech recognition. In *Proceedings of the 97th Meeting of the Acoustical Society of America*, pages 547–550.

Michael Brent. 1993. Minimal generative explanations: A middle ground between neurons and triggers. In *Proc. of the 15th Annual Meeting of the Cognitive Science Society*, pages 28–36.

Eric Brill. 1993. Automatic grammar induction and parsing free text: A transformation based approach. In *Proceedings of the DARPA Speech and Natural Language Workshop*.

Ted Briscoe and Nick Waegner. 1992. Robust stochastic parsing using the inside-outside algorithm. In *Proc. of the AAAI Workshop on Probabilistic-Based Natural Language Processing Techniques*, pages 39–52.

Timothy Andrew Cartwright and Michael R. Brent. 1994. Segmenting speech without a lexicon: Evidence for a bootstrapping model of lexical acquisition. In *Proc. of the 16th Annual Meeting of the Cognitive Science Society*, Hillsdale, New Jersey.

Carl de Marcken. 1995. Parsing with stochastic, feature-based RTNs. Memo A.I. Memo, MIT Artificial Intelligence Lab., Cambridge, Massachusetts.

S. Della-Pietra, V. Della-Pietra, J. Gillett, J. Lafferty, H. Printz, and L. Ureš. 1994. Inference and estimation of a long-range trigram model. In *International Colloquium on Grammatical Inference*, pages 78–92, Alicante, Spain.

T. Mark Ellison. 1992. *The Machine Learning of Phonological Structure.* Ph.D. thesis, University of Western Australia.

John Lafferty, Daniel Sleator, and Davy Temperley. 1992. Grammatical trigrams: A probabalistic model of link grammar. Technical Report CMU-CS-92-181, Carnegie Mellon University, Pittsburgh, Pennsylvania.

K. Lari and S. J. Young. 1990. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language*, 4:35–56.

David M. Magerman and Mitchell P. Marcus. 1990. Parsing a natural language using mutual information statistics. In *Proc. of the American Association for Artificial Intelligence*, pages 984–989.

Mitchell Marcus. 1991. Very large annotated databse of American English. In *Proceedings of the DARPA Speech and Natural Language Workshop.*

I. A. Melčuk. 1988. *Dependency Syntax: Theory and Practice.* State University of New York Press.

Donald Cort Olivier. 1968. *Stochastic Grammars and Language Acquisition Mechanisms.* Ph.D. thesis, Harvard University, Cambridge, Massachusetts.

Fernando Pereira and Yves Schabes. 1992. Inside-outside reestimation from partially bracketed corpora. In *Proc. 29th Annual Meeting of the Association for Computational Linguistics,* pages 128–135, Berkeley, California.

Daniel D. K. Sleator and Davy Temperley. 1991. Parsing english with a link grammar. Technical Report CMU-CS-91-196, Carnegie Mellon University, Pittsburgh, Pennsylvania.

Andreas Stolcke. 1994. *Bayesian Learning of Probabalistic Language Models.* Ph.D. thesis, University of California at Berkeley, Berkeley, CA.

J. Gerald Wolff. 1982. Language acquisition, data compression and generalization. *Language and Communication*, 2(1):57–89.