PROBABILISTIC TREE-ADJOINING GRAMMAR AS A FRAMEWORK FOR STATISTICAL NATURAL LANGUAGE PROCESSING

Philip Resnik*
Department of Computer and Information Science
University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA
resnik@linc.cis.upenn.edu

Abstract

In this paper, I argue for the use of a probabilistic form of tree-adjoining grammar (TAG) in statistical natural language processing. I first discuss two previous statistical approaches — one that concentrates on the probabilities of structural operations, and another that emphasizes co-occurrence relationships between words. I argue that a purely structural approach, exemplified by probabilistic context-free grammar, lacks sufficient sensitivity to lexical context, and, conversely, that lexical co-occurence analyses require a richer notion of locality that is best provided by importing some notion of structure.

I then propose probabilistic TAG as a framework for statistical language modelling, arguing that it provides an advantageous combination of structure, locality, and lexical sensitivity. Issues in the acquisition of probabilistic TAG and parameter estimation are briefly considered.

(S_1)	S	→	NP	VP	(0.5)
(S_2)	S	→	S	PP	(0.35)
(S_3)	S	→	V	NP	(0.15)
(VP_1)	VP	→	V	NP	(0.6)
(VP_2)	VP	→	V	PP	(0.4)
	⋮				

Figure 1: *Fragment of a context-free grammar.*

S	
NP VP	(r_1 = S → NP VP)
N VP	(r_2 = NP → N)
we VP	(r_3 = N → we)
we V NP	(r_4 = VP → V NP)
we like NP	(r_5 = V → like)
we like N	(r_6 = NP → N)
we like Mary	(r_7 = N → Mary)

Figure 2: *A context-free derivation*

1 Probabilistic CFGs

Thanks to the increased availability of text corpora, fast computers, and inexpensive off-line storage, statistical approaches to issues in natural language processing are enjoying a surge of interest, across a wide variety of applications. As research in this area progresses, the question of how to combine our existing knowledge of grammatical methods (e.g. generative power, efficient parsing) with developments in statistical and information-theoretic methods (especially techniques imported from the speech-processing community) takes on increasing significance.

Perhaps the most straightforward combination of grammatical and statistical techniques can be found in the probabilistic generalization of context-free grammars. [Jelinek *et al.*, 1990] characterize a probabilistic context-free grammar (PCFG) as a context-free grammar in which each production has been assigned a probability of use. I will refer to the entire set of probabilities as the *statistical parameters*, or simply *parameters*, of the probabilistic grammar.

For example, the productions in Figure 1, together with their associated parameters, might comprise a fragment of a PCFG for English. Note that for each nonterminal symbol, the probabilities of its alternative expansions sum to 1. This captures the fact that in a context-free derivation, each instance of a nonterminal symbol must be rewritten according to exactly one of its expansions. Also by definition of a context-free derivation, each rewriting is independent of the context within which the nonterminal appears. So, for example, in the (leftmost) derivation of *We like Mary* (Figure 2), the expansions N → *we* and VP → V NP are independent events.

The probability of n independent events is

*This research was supported by the following grants: ARO DAAL 03-89-C-0031, DARPA N00014-90-J-1863, NSF IRI 90-16592, and Ben Franklin 91S.3078C-1. I would like to thank Aravind Joshi, Yves Schabes, and members of the CLiFF group at Penn for helpful discussion.

merely the product of the probabilities of each individual event. Therefore the probability of a context-free derivation with rewritings $r_1, r_2, \ldots, r_n$ is $\Pr(r_1)\Pr(r_2)\cdots\Pr(r_n)$.

Jelinek *et al.* use this fact to develop a set of efficient algorithms (based on the CKY parsing algorithm) to compute the probability of a sentence given a grammar, to find the most probable parse of a given sentence, to compute the probability of an initial substring leading to a sentence generated by the grammar, and, following [Baker, 1979], to estimate the statistical parameters for a given grammar using a large text corpus.

Now, the definition of the probability of a derivation that is used for PCFG depends crucially upon the context-freeness of the grammar; without it, the independence assumption that permits us to simply multiply probabilities is compromised. Yet experience tells us that rule expansions are *not*, in general, context-free. Consider the following simple example. According to one frequency estimate of English usage [Francis and Kucera, 1982], *he* is more than three times as likely as *we* to appear as a subject pronoun. So if *we* is replaced with *he* in Figure 2, the new derivation (of *He like Mary*) is accorded far higher probability according to the PCFG, even though *He like Mary* is not English. The problem extends beyond such obvious cases of agreement: since *crushing* happens to be more likely than *mashing*, the probability of deriving *He's crushing potatoes* is greater than the probability corresponding derivation for *He's mashing potatoes*, although we expect the latter to be more likely.

In short, lexical context matters. Although probabilistic context-free grammar captures the fact that not all nonterminal rewritings are equally likely, its insensitivity to lexical context makes it less than adequate as a statistical model of natural language,[1] and weakens Jelinek *et al.*'s contention that "in an ambiguous but appropriately chosen probabilistic CFG (PCFG), correct parses are high probability parses" (p. 2). In practical terms, it also suggests that techniques for incorporating lexical context-sensitivity will potentially improve PCFG performance.

2 Co-occurrence Statistics

A second common approach to the corpus-based investigation of natural language considers instances of words occurring adjacent to each other in the text, or, more generally, words occurring together within a window of fixed size. For example, [Church and Hanks, 1989] use an information-theoretic measure, mutual information, to calculate the degree of association of word pairs based upon their co-occurrence throughout a large corpus. The mutual information between two events is defined as

$$\mathrm{I}(x;y) = \log\frac{\Pr(x,y)}{\Pr(x)\Pr(y)}$$

where, for the purposes of corpus analysis, $\Pr(x)$ and $\Pr(y)$ are the respective probabilities of words x and y appearing within the corpus, and $\Pr(x,y)$ is the probability of word x being followed by word y within a window of w words. Intuitively, mutual information relates the actual probability of seeing x and y together (numerator) to the probability of seeing them together under the assumption that they were independent events (denominator).

As defined here, the calculation of mutual information takes into account only information about co-occurrence within surface strings. A difficulty with such an approach, however, is that co-occurrences of interest may not be sufficiently local. Although sentence (1)a can witness the co-occurence of *students* and *spend* within a small window (say, $w = 2$ or 3), sentence (1)b cannot.

(1)a. Students spend a lot of money.
b. Students who attend conferences spend a lot of money.

Simply increasing the window size will not suffice, for two reasons. First, there is no bound on the length of relative clauses such as the one in (1)b, hence no fixed value of w that will solve the problem. Second, the choice of window size depends on the application — Church and Hanks write that "smaller window sizes will identify fixed expressions (idioms) and other relations that hold over short ranges; larger window sizes will highlight semantic concepts and other relationships that hold over larger scales" (p. 77). Extending the window size in order to capture co-occurrences such as the one found in (1)b may therefore undermine other goals of the analysis.

[Brown *et al.*, 1991] encounter a similar problem. Their application is statistics-driven machine translation, where the statistics are calculated on the basis of trigrams (observed triples) of words within the source and target corpora. As in (1), sentences (2)a and (3)a illustrate a difficulty encountered when limiting window size. Light verb constructions are an example of a situation in which the correct translation of the verb depends on the identity of its object. The correct word sense of *prendre*, "to make," is chosen in (2)a, since the dependency signalling that word sense — between *prendre* and its object, *décision* — falls within a trigram window and thus within the bounds of their language model.

(2)a. *prendre la décision*

[1] Independent of the statistical issue, of course, it is also generally accepted that CFGs are generatively too weak to model natural language in its full generality [Shieber, 1985].

b. make the decision

(3)a. *prendre une difficile décision*

b. *take a difficult decision

In (3)a, however, the dependency is not sufficiently local: the model has no access to lexical relationships spanning a distance of more than three words. Without additional information about the verb's object, the system relies on the fact that the sense of *prendre* as "to take" occurs with greater frequency, and the incorrect translation results.

Brown *et al.* propose to solve the problem by seeking clues to a word's sense within a larger context. Each possible clue, or *informant*, is a site relative to the word. For *prendre*, the informant of interest is "the first noun to the right"; other potential informants include "the first verb to the right," and so on. A set of such potential informants is defined in advance, and statistical techniques are used to identify, for each word, which potential informant contributes most to determining the sense of the word to use.

It seems clear that in most cases, Brown *et al.*'s informants represent approximations to structural relationships such as subject and object. Examples (1) through (3) suggest that attention to structure would have advantages for the analysis of lexical relationships. By using co-occurrence within a structure rather than co-occurrence within a window, it is possible to capture lexical relationships without regard to irrelevant intervening material. One way to express this notion is by saying that structural relationships permit us to use a notion of *locality* that is more general than simple distance in the surface string. So, for example, (1) demonstrates that the relationship between a verb and its subject is not affected by a subject relative clause of any length.

3 Hybrid Approaches

In the previous sections I have observed that probabilistic CFGs, though capable of capturing the relative likelihoods of context-free derivations, require greater sensitivity to lexical context in order to model natural languages. Conversely, lexical co-occurrence analyses seem to require a richer notion of locality, something that structural relationships such as subject and object can provide. In this section I briefly discuss two proposals that could be considered "hybrid" approaches, making use of both grammatical structure and lexical co-occurrence.

The first of these, [Hindle, 1990], continues in the direction suggested at the end of the previous section: rather than using "informants" to approximate structural relationships, lexical co-occurrence is calculated over structures. Hindle uses co-occurrence statistics, collected over parse trees, in order to classify nouns on the basis of the verb contexts in which they appear. First, a robust parser is used to obtain a parse tree (possibly partial) for each sentence. For example, the following table contains some of the information Hindle's parser retrieved from the parse of the sentence "The clothes we wear, the food we eat, the air we breathe, the water we drink, the land that sustains us, and many of the products we use are the result of agricultural research."

verb	subject	object
eat	we	food
breathe	we	air
drink	we	water
sustain	land	us

Next, Hindle calculates a similarity measure based on the mutual information of verbs and their arguments: nouns that tend to appear as subjects and objects of the same verbs are judged more similar. According to the criterion that that two nouns be reciprocally most similar to each other, Hindle's analysis derives such plausible pairs of similar nouns as *ruling* and *decision*, *battle* and *fight*, *researcher* and *scientist*.

A proposal in [Magerman and Marcus, 1991] relates to probabilistic parsing. As in PCFG, Magerman and Marcus associate probabilities with the rules of a context-free grammar, but these probabilities are *conditioned* on the contexts within which a rule is observed to be used. It is in the formulation of "context" that lexical co-occurrence plays a role, albeit an indirect one.

Magerman and Marcus's Pearl parser is essentially a bottom-up chart parser with Earley-style top-down prediction. When a context-free rule $A \rightarrow \alpha_1 \ldots \alpha_k$ is proposed as an entry in the chart spanning input symbols a_i through a_j, it is assigned a "score" based on

1. the rule that generated this instance of nonterminal symbol A, and
2. the part-of-speech trigram centered at a_i.

For example, given the input *My first love was named Pearl*, a proposed chart entry VP → V NP starting its span at *love* (i.e., a theory trying to interpret *love* as a verb) would be scored on the basis of the rule that generated the VP (in this case, probably S → NP VP) together with the part-of-speech trigram "adjective verb verb." This score would be lower than that of a different chart entry interpreting *love* as a noun, since the latter would be scored using the more likely part of speech trigram "adjective noun verb". So, although the context-free probability favors the interpretation of *love* as the beginning of a verb phrase,

information about the lexical context of the word rescues the correct interpretation, in which it is categorized as a noun.

Although Pearl does not take into account the relationships between particular words, it does represent a promising combination of CFG-based probabilistic parsing and corpus statistics calculated on the basis of simple (trigram) co-occurrences.

A difficulty with the hybrid approaches, however, is that they leave unclear exactly what the statistical model of the language *is*. In probabilistic context-free grammar and in surface-string analysis of lexical co-occurrence, there is a precise definition of the *event space* — what events go into making up a sentence. (As discussed earlier, the events in a PCFG derivation are the rule expansions; the events in a window-based analysis of surface strings can be viewed as transitions in a finite Markov chain modelling the language.) The absence of such a characterization in the hybrid approaches makes it more difficult to identify what assumptions are being made, and gives such work a decidedly empirical flavor.

4 Probabilistic TAG

4.1 Lexicalized TAG

The previous sections have demonstrated a need for a well-defined statistical framework in which both syntactic structure and lexical co-occurrences are incorporated. In this section, I argue that a probabilistic form of lexicalized tree-adjoining grammar provides just such a seamless combination. I begin with a brief description of lexicalized tree-adjoining grammar, and then present its probabilistic generalization and the advantages of the resulting formalism.

Tree-adjoining grammar [Joshi *et al.*, 1975], or TAG, is a generalization of context-free grammar that has been proposed as a useful formalism for the study of natural languages. A tree-adjoining grammar comprises two sets of elementary structures: initial trees and auxiliary trees. (See Figure 3, in which initial and auxiliary trees are labelled using α and β, respectively.) An auxiliary tree, by definition, has a nonterminal node on its frontier — the *foot* node — that matches the nonterminal symbol at its root.

These elementary structures can be combined using two operations, *substitution* and *adjunction*. The substitution operation corresponds to the rewriting of a symbol in a context-free derivation: a nonterminal node at the frontier of a tree is replaced with an initial tree having the same nonterminal symbol at its root. So, for example, one could expand either NP in α_1 by rewriting it as tree α_2.

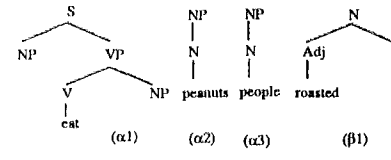

Figure 3: *Examples of TAG elementary trees*

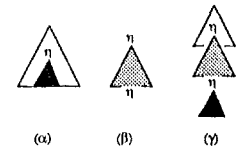

Figure 4: *Adjunction of auxiliary tree β into tree α at node η, resulting in derived structure γ.*

The adjunction operation is a generalization of substitution that permits internal as well as frontier nodes to be expanded. One adjoins an auxiliary tree β into a tree α at node η by "splitting" α horizontally at η and then inserting β. In the resulting structure, γ, β's root node appears in η's original position, and its foot node dominates all the material that η dominated previously (see Figure 4). For example, Figure 5 shows γ_1, the result of adjoining auxiliary tree β_1 at the N node of α_2.

For context-free grammar, each derived structure is itself a record of the operations that produced it: one can simply read from a parse tree the context-free rules that were used in the parse. In contrast, a derived structure in a tree-adjoining grammar is distinct from its derivation history. The final parse tree encodes the *structure* of the sentence according to the grammar, but the *events* that constitute the derivation history — that is, the subsitutions and adjunctions that took place — are not directly encoded. The significance of this distinction will become apparent shortly.

A *lexicalized* tree-adjoining grammar is a TAG in which each elementary structure (initial or auxiliary tree) has a lexical item on its frontier, known as its *anchor* [Schabes, 1990]. Another natural way to say this is that each lexical item has associated with it

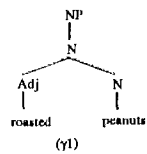

Figure 5: *Result of adjoining β_1 into α_2.*

a set of structures that, taken together, characterize the contexts within which that item can appear.[2] For example, tree α_1 in Figure 3 represents one possible structural context for *eat*; another would be an initial tree in which *eat* appears as an intransitive verb, and yet another would be a tree containing *eat* within a passive construction.

4.2 Features of Probabilistic TAG

Three features of lexicalized TAG make it particularly appropriate as a probabilistic framework for natural language processing. First, since every tree is associated with a lexical anchor, words and their associated structures are tightly linked. Thus, unlike probabilistic context-free grammar, the probabilities associated with structural operations are sensitive to lexical context. In particular, the probability of an adjunction step such as the one that produced γ_1, above, is sensitive to the lexical anchors of the trees. One would expect a similar adjunction of β_1 into α_3 (resulting in the string *roasted people*) to have extremely low probability, reflecting the low degree of association between the lexical items involved. Similarly, tree α_3 is far more likely than α_2 to be substituted as the subject NP in tree α_1, since *people* is more likely than *peanuts* to appear as the subject of *eat*.

Notice that this attention to lexical context is not acquired at the expense of the independence assumption for probabilities. Just as expansion of a nonterminal node in a CFG derivation takes place without regard to that node's history, substitution or adjunction to a node during a TAG derivation takes place without regard to that node's history.[3] Thus it will be straightforward to express the probability of a probabilistic TAG derivation as a product of probabilities, just as was the case for PCFG, yielding a well-defined statistical model.

Second, since TAG factors local dependencies from recursion, the problem of intervening material in window-based lexical co-occurrence analyses does not arise. Notice that in tree α_1, the verb, its subject, and its object all appear together within a single elementary structure. The introduction of recursive substructure — relative clauses, adverbs, strings of adjectival modifiers — is done entirely by means of the adjunction operation, as illustrated by figures 4 and 5. This fact provides a principled treatment of examples such as (4): the probability of substituting α_2 as the object NP of α_1 (capturing an association between *eat* and *peanuts*) is independent of any other operations that might take place, such as the introduction (via adjunction) of an adjectival modifier. Similarly, the probability of substituting α_3 as the subject NP of α_1 is not affected by the subsequent adjunction of a relative clause (cf. example (1)). Thus, in contrast to a co-occurrence analysis based on strings, the analysis of (4) within probabilistic TAG finds precisely the same relationship between the verb and its arguments in (4)b and (4)c as it does in (4)a.

(4)a. People eat peanuts.
b. People eat roasted peanuts.
c. People who are visiting the zoo eat roasted peanuts.

Third, the notion of lexical anchor in lexicalized TAG has been generalized to account for multi-word lexical entries [Abeillé and Schabes, 1989]. Thus the formalism appears to satisfy the criteria of [Smadja and McKeown, 1990], who write of the need for "a flexible lexicon capable of using single word entries, multiple word entries as well as phrasal templates and a mechanism that would be able to gracefully merge and combine them with other types of constraints" (p. 256). Among "other types of constraints" are linguistic criteria, and here, too, TAG offers the potential for capturing the details of linguistic and statistical facts in a uniform way [Kroch and Joshi, 1985].

4.3 Formalization of Probabilistic TAG

There are a number of ways to express lexicalized tree-adjoining grammar as a probabilistic grammar formalism.[4] Here I propose what appears to me to be the most direct probabilistic generalization of lexicalized TAG; a different treatment can be found in [Schabes, 1992].

Definitions:

Let I denote the set of initial trees in the grammar, and A the set of auxiliary trees.

Each tree in a lexicalized TAG has a (possibly empty) subset of its frontier nodes marked as nodes at which substitution may take place. Given a tree α, let that subset be denoted by $s(\alpha)$.

Adjunction may take place at any node η in α labelled by a nonterminal symbol, so long as $\eta \notin s(\alpha)$. Denote this set of possible adjunction nodes $a(\alpha)$.[5]

[2] Lexicalized TAG is only one of many such grammar formalisms, of course.

[3] Schabes (personal communication) captures the generalization quite nicely in observing that for both CFG derivation trees and TAG derivation histories, the path sets (set of possible paths from root to frontier) are regular.

[4] The idea of defining probabilities over derivations involving combinations of elementary structures was introduced as early as [Joshi, 1973].

[5] Note that substitution nodes and adjunction nodes are not distinguished in Figure 3.

Let $S(\alpha, \alpha', \eta)$ denote the event of substituting tree α' into tree α at node η.

Let $A(\alpha, \beta, \eta)$ denote the event of adjoining auxiliary tree β into tree α at node η, and let $A(\alpha, none, \eta)$ denote the event in which no adjunction is performed at that node.

Let Ω = the set of all substitution and adjunction events.

A *probabilistic tree-adjoining grammar* is a 5-tuple, $\langle I, A, P_I, P_S, P_A \rangle$, where I and A are defined as above, P_I is a function from I to the real interval [0,1], and P_S and P_A are functions from Ω to [0,1], such that:

1. $$\sum_{\alpha \in I} P_I(\alpha) = 1$$

2. $$\forall_{\alpha \in I \cup A} \forall_{\eta \in s(\alpha)} \sum_{\alpha' \in I} P_S(S(\alpha, \alpha', \eta)) = 1$$

3. $$\forall_{\alpha \in I \cup A} \forall_{\eta \in a(\alpha)} \sum_{\beta \in A \cup \{none\}} P_A(A(\alpha, \beta, \eta)) = 1$$

$P_I(\alpha)$ is interpreted as the probability that a derivation begins with initial tree α. $P_S(S(\alpha, \alpha', \eta))$ denotes the probability of substituting α' at node η of tree α, and $P_A(A(\alpha, \beta, \eta))$ denotes the probability of adjoining β at node η of tree α (where $P_A(A(\alpha, none, \eta))$ denotes the probability that no adjunction takes place at node η of α).

A TAG derivation is described by the initial tree with which it starts, together with the sequence of substitutions and adjunction operations that then take place. Denoting each operation as $op(\alpha_1, \alpha_2, \eta), op \in \{S, A\}$, and denoting the initial tree with which the derivation starts as α_0, the *probability of a TAG derivation* $\tau = \langle \alpha_0, op_1(\cdots), \ldots, op_n(\cdots) \rangle$ is

$$\Pr(\tau) = P_I(\alpha_0) \prod_{1 \le i \le n} P_{op_i}(op_i(\cdots))$$

This definition is directly analogous to the probability of a context-free derivation $\tau = \langle r_1, \ldots, r_n \rangle$,

$$\Pr(\tau) = P_I(S) \prod_{1 \le i \le n} P(r_i),$$

though in a CFG every derivation starts with S and so $P_I(S) = 1$.

Does probabilistic TAG, thus formalized, behave as desired? Returning to example (4), consider the following derivation history for the sentence *People eat roasted peanuts*:

$\langle \alpha_1,$
$S(\alpha_1, \alpha_3, \mathrm{NP}_{subj}),$
$S(\alpha_1, \alpha_2, \mathrm{NP}_{obj}),$
$A(\alpha_2, \beta_1, \mathrm{N}) \rangle$

Notice first that, given intuitively reasonable estimates for the probabilities of substitution and adjunction, the probability of this derivation would indeed be far greater than for the corresponding derivation of *Peanuts eat roasted people*. Thus, in contrast to PCFG, probabilistic TAG's sensitivity to lexical context does provide a more accurate language model. In addition, were this derivation to be used as an observation for the purpose of *estimating* probabilities, the estimate of $P_S(S(\alpha_1, \alpha_3, \mathrm{NP}_{subj}))$ would be unaffected by the event $A(\alpha_2, \beta_1, \mathrm{N})$. That is, the model would in fact capture a relationship between the verb *eat* and its object *peanuts*, mediated by the trees that they anchor, regardless of intervening material in the surface string.

4.4 Acquisition of Probabilistic TAG

Despite the attractive theoretical features of probabilistic TAG, many practical issues remain. Foremost among these is the acquisition of the statistical parameters. [Schabes, 1992] has recently adapted the Inside-Outside algorithm, used for estimating the parameters of a probabilistic CFG [Baker, 1979], to probabilistic TAG. The Inside-Outside algorithm is itself a generalization of the Forward-Backward algorithm used to train hidden Markov models [Baum, 1972]. It optimizes by starting with a set of initial parameters (chosen randomly, or uniformly, or perhaps initially set on the basis of *a priori* knowledge), then iteratively collecting statistics over the training data (using the existing parameters) and then re-estimating the parameters on the basis of those statistics. Each re-estimation step guarantees improved or at worst equivalent parameters, according to a maximum-likelihood criterion.

The procedure immediately raises two questions, regardless of whether the probabilistic formalism under consideration is finite-state, context-free, or tree-adjoining. First, since the algorithm re-estimates parameters but does not determine the rules in the grammar, the structures underlying the statistics (arcs, rules, trees) must be determined in some other fashion. The starting point can be a hand-written grammar, which requires considerable effort and linguistic knowledge; alternatively, one can initially include all possible structures and let the statistics weed out useless rules by assigning them near-zero probability. A third possibility is to add an engine that hypothesizes rules on the basis of observed data, and then let the parameter estimation operate over these hypothesized structures. Which possibility is best depends upon the application and the breadth of linguistic coverage needed.

Second, any statistical approach to natural language must consider the size of the parameter set to be estimated. Additional parameters can enhance the theoretical accuracy of a statistical model (e.g., extending a trigram model to 4-grams) but may also lead to an

unmanageable number of parameters to store and retrieve, much less estimate. In addition, as the number of parameters grows, more data are required in order to collect accurate statistics. For example, both (5)a and (5)b witness the same relationship between *called* and *son* within a trigram model.

(5)a. Mary called her son.
b. Mary called her son Junior.

Within a recent lexicalized TAG for English [Abeillé *et al.*, 1990], however, these two instances of *called* are associated with two distinct initial trees, reflecting the different syntactic structures of the two examples. Thus observations that would be the same for the trigram model can be fragmented in the TAG model. As a result of this fragmentation, each individual event is observed fewer times, and so the model is more vulnerable to statistical inaccuracy resulting from low counts.

The acquisition of a probabilistic tree-adjoining grammar "from the ground up," that is, including hypothesizing grammatical structures as well as estimating statistical parameters, is an intended topic of future work.

5 Conclusions

I have argued that probabilistic TAG provides a seamless framework that combines the simplicity and structure of the probabilistic CFG approach with the lexical sensitivity of string-based co-occurrence analyses. Within a wide variety of applications, it appears that various researchers (e.g. [Hindle, 1990; Magerman and Marcus, 1991; Brown *et al.*, 1991; Smadja and McKeown, 1990]) are confronting issues similar to those discussed here. As the resources required for statistical approaches to natural language continue to become more readily available, these issues will take on increasing importance, and the need for a framework that unifies grammatical and statistical techniques will continue to grow. Probabilistic TAG is one step toward such a unifying framework.

References

[Abeillé and Schabes, 1989] Anne Abeillé and Yves Schabes. Parsing idioms in tree adjoining grammars. In *Fourth Conference of the European Chapter of the Association for Computational Linguistics (EACL'89)*, Manchester, 1989.

[Abeillé *et al.*, 1990] Anne Abeillé, Kathleen Bishop, Sharon Cote, and Yves Schabes. A lexicalized tree adjoining grammar for english. Technical Report MS-CIS-90-24, Department of Computer and Information Science, University of Pennsylvania, April 1990.

[Baker, 1979] J.K. Baker. Trainable grammars for speech recognition. In *Proceedings of the Spring Conference of the Acoustical Society of America*, pages 547–550, Boston, MA, June 1979.

[Baum, 1972] L.E. Baum. An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process. *Inequalities*, 3:1–8, 1972.

[Brown *et al.*, 1991] P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. A statistical approach to sense disambiguation in machine translation. In *Fourth DARPA Workshop on Speech and Natural Language*, Pacific Grove, CA, February 1991.

[Church and Hanks, 1989] K. Church and P. Hanks. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th Meeting of the Association for Computational Linguistics*, 1989. Vancouver, B.C.

[Francis and Kucera, 1982] W. Francis and H. Kucera. *Frequency Analysis of English Usage.* Houghton Mifflin Co.: New York, 1982.

[Hindle, 1990] D. Hindle. Noun classification from predicate-argument structures. In *Proceedings of the 28th Annual Meeting of the Assocation of Computational Linguistics, Pittsburgh, Penna.*, pages 268–275, 1990.

[Jelinek *et al.*, 1990] F. Jelinek, J. D. Lafferty, and R. L. Mercer. Basic methods of probabilistic context free grammars. Research Report RC 16374 (#72684), IBM, Yorktown Heights, New York 10598, 1990.

[Joshi *et al.*, 1975] Aravind Joshi, Leon Levy, and M Takahashi. Tree adjunct grammars. *Journal of the Computer and System Sciences*, 10, 1975.

[Joshi, 1973] Aravind Joshi. Remarks on some aspects of language structure and their relevance to pattern analysis. *Pattern Recognition*, 5:365–381, 1973.

[Kroch and Joshi, 1985] Anthony Kroch and Aravind K. Joshi. Linguistic relevance of tree adjoining grammars. Technical Report MS-CIS-85-18, Department of Computer and Information Science, University of Pennsylvania, April 1985.

[Magerman and Marcus, 1991] D. Magerman and M. Marcus. Pearl: A probabilistic chart parser. In *Fourth DARPA Workshop on Speech and Natural Language*, Pacific Grove, CA, February 1991.

[Schabes, 1990] Yves Schabes. *Mathematical and Computational Aspects of Lexicalized Grammars.* PhD thesis, Univ. of Pennsylvania, 1990.

[Schabes, 1992] Yves Schabes. Stochastic lexicalized tree-adjoining grammars, 1992. This proceedings.

[Shieber, 1985] Stuart Shieber. Evidence against the context-freeness of natural language. *Linguistics and Philosophy*, 8:333–343, 1985.

[Smadja and McKeown, 1990] F. Smadja and K. McKeown. Automatically extracting and representing collocations for language generation. In *Proceedings of the 28th Annual Meeting of the Assocation of Computational Linguistics*, pages 252–259, 1990.