A STOCHASTIC APPROACH TO SENTENCE PARSING

Tetsunosuke Fujisaki
Science Institute, IBM Japan, Ltd.
No. 36 Kowa Building
5-19 Sanbancho,Chiyoda-ku
Tokyo 102, Japan

ABSTRACT

A description will be given of a procedure to assign the most likely probabilities to each of the rules of a given context-free grammar. The grammar developed by S. Kuno at Harvard University was picked as the basis and was successfully augmented with rule probabilities. A brief exposition of the method with some preliminary results, when used as a device for disambiguating parsing English texts picked from natural corpus, will be given.

I. INTRODUCTION

To prepare a grammar which can parse arbitrary sentences taken from a natural corpus is a difficult task. One of the most serious problems is the potentially unbounded number of ambiguities. Pure syntactic analysis with an imprudent grammar will sometimes result in hundreds of parses.

With prepositional phrase attachments and conjunctions, for example, it is known that the actual growth of ambiguities can be approximated by a Catlan number [Knuth], the number of ways to insert parentheses into a formula of N terms: 1, 2, 5, 14, 42, 132, 469, 1430, 4892, ... The five ambiguities in the following sentence with three ambiguous constructions can be well explained with this number.

> I saw a man in a park with a scope.

This Catalan number is essentially exponential and [Martin] reported a syntactically ambiguous sentence with 455 parses:

> List the sales of products produced in 1973
> with the products produced in 1972.

On the other hand, throughout the long history of natural language understanding work, semantic and pragmatic constraints are known to be indispensable and are recommended to be represented in some formal way and to be referred to during or after the syntactic analysis process.

However, to represent semantic and pragmatic constraints, (which are usually domain sensitive) in a well-formed way is a very difficult and expensive task. A lot of effort in that direction has been expended, especially in Artificial Intelligence, using semantic networks, frame theory, etc. However, to our knowledge no one has ever succeeded in preparing them except in relatively small restricted domains. [Winograd, Sibuya].

Faced with this situation, we propose in this paper to use statistics as a device for reducing ambiguities. In other words, we propose a scheme for grammatical inference as defined by [Fu], a stochastic augmentation of a given grammar; furthermore, we propose to use the resultant statistics as a device for semantic and pragmatic constraints. Within this stochastic framework, semantic and pragmatic constraints are expected to be coded implicitly in the statistics. A simple bottom-up parse referring to the grammar rules as well as the statistics will assign relative probabilities among ambiguous derivations. And these relative probabilities should be useful for filtering meaningless garbage parses because high probabilities will be assigned to the parse trees corresponding to meaningful interpretations and low probabilities, hopefully 0.0, to other parse trees which are grammatically correct but are not meaningful.

Most importantly, stochastic augmentation of a grammar will be done automatically by feeding a set of sentences as samples from the relevant domain in which we are interested, while the preparation of semantic and pragmatic constraints in the form of usual semantic network, for example, should be done by human experts for each specific domain.

This paper first introduces the basic ideas of automatic training process of statistics from given example sentences, and then shows how it works wit experimental results.

II. GRAMMATICAL INFERENCE OF A STOCHASTIC GRAMMAR

A. Estimation of Markov Parameters for sample texts

Assume a Markov source model as a collection of states connected to one another by transitions which produce symbols from a finite alphabet. To each transition, t from a state s, is associated a probability $q(s,t)$, which is the probability that t will be chosen next when s is reached.

When output sentences $\{B(i)\}$ from this markov model are observed, we can estimate the transition probabilities $\{q(s,t)\}$ through an iteration process in the following way:

1. Make an initial guess of $\{q(s,t)\}$.

2. Parse each output sentence $B(i)$. Let $d(i,j)$ be a j-th derivation of the i-th output sentence $B(i)$.

3. Then the probability $p(d(i,j))$ of each derivation $d(i,j)$ can be defined in the following way:

   $p(d(i,j))$ is the product of probability of all the transitions $q(s,t)$ which contribute to that derivation $d(i,j)$.

4. From this $p(d(i,j))$, the Bayes a posteriori estimate of the count $c(s,t,i,j)$, how many times the transition t from state s is used on the derivation $d(i,j)$, can be estimated as follows:

$$c(s,t,i,j) = \frac{n(s,t,i,j) \times p(d(i,j))}{\sum_j p(d(i,j))}$$

   where $n(s,t,i,j)$ is a number of times the transition t from state s is used in the derivation $d(i,j)$.

   Obviously, $c(s,t,i,j)$ becomes $n(s,t,i,j)$ in an unambiguous case.

5. From this $c(s,t,i,j)$, new estimate of the probabilities $f(s,t)$ can be calculated.

$$f(s,t) = \frac{\sum_i \sum_j c(s,t,i,j)}{\sum_i \sum_j \sum_t c(s,t,i,j)}$$

6. Replace $\{q(s,t)\}$ with this new estimate $\{f(s,t)\}$ and repeat from step 2.

Through this process, asymptotic convergence will hold in the entropy of $\{q(s,t)\}$ which is defined as:

$$Entoropy = \sum_s \sum_t -q(s,t) \times \log(q(s,t))$$

and the $\{q(s,t)\}$ will approach the real transition probability [Baum-1970,1792].

Further optimized versions of this algorithm can be found in [Bahl-1983] and have been successfully used for estimating parameters of various Markov models which approximate speech processes [Bahl - 1978, 1980].

B. Extension to context-free grammar

This procedure for automatically estimating Markov source parameters can easily be extended to context-free grammars in the following manner.

Assume that each state in the Markov model corresponds to a possible sentential form based on a given context-free grammar. Then each transition corresponds to the application of a context-free production rule to the previous state, i.e. previous sentential form. For example, the state NP.VP can be reached from the state S by applying a rule S->NP VP, the state ART.NOUN.VP can be reached from the state NP.VP by applying the rule NP->ART NOUN to the first NP of the state NP.VP, and so on.

Since the derivations correspond to sequences of state transitions among the states defined above, parsing over the set of sentences given as training data will enable us to count how many times each transition is fired from the given sample sentences. For example, transitions from the state S to the state NP.VP may occur for almost every sentence because the corresponding rule, 'S->NP VP', must be used to derive the most frequent declarative sentences; the transition from state ART.NOUN.VP to the state 'every'.NOUN.VP may happen 103 times; etc. If we associate each grammar rule with an a priori probability as an initial guess, then the Bayes a posteriori estimate of the number of times each transition will be traversed can be calculated from the initial probabilities and the actual counts observed as described above.

Since each production is expected to occur independently of the context, the new estimate of the probability for a rule will be calculated at each iteration step by masking the contexts. That is, the Bayes estimate counts from all of the transitions which correspond to a single context free rule; all transitions between states like xxx.A.yyy and xxx.B.C.yyy correspond to the production rule 'A->B C' regardless of the contents of xxx and yyy; are tied together to get the new probability estimate of the corresponding rule.

Renewing the probabilities of the rules with new estimates, the same steps will be repeated until they converge.

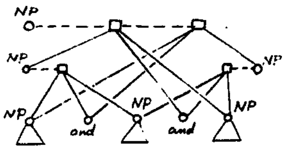
III. EXPERIMENTATION

A. Base Grammar

As the basis of this research, the grammar developed by Prof. S. Kuno in the 1960's for the machine translation project at Harvard University [Kuno-1963, 1966] was chosen, with few modifications. The set of grammar specifications in that grammar, which are in Greibach normal form, were translated into a form which is favorable to our method. 2118 rules of the original rules were rewritten as 5241 rules in Chomsky normal form.

B. Parser

A bottom-up context-free parser based on Cocke-Kasami-Young algorithm was developed especially for this purpose. Special emphasis was put on the design of the parser to get better performance in highly ambiguous cases. That is, alternative-links, the dotted link shown in the figure below, are introduced to reduce the number of intermediate substructure as far as possible.



C. Test Corpus

Training sentences were selected from the magazines, 31 articles from Reader's Digest and Datamation, and from IBM correspondence. Among 5528 selected sentences from the magazine articles, 3582 sentences were successfuly parsed with 0.89 seconds of CPU time (IBM 3033-UP) and with 48.5 ambiguities per a sentence. The average word lengths were 10.85 words from this corpus.

From the corpus of IBM correspondence, 1001 sentences, 12.65 words in length in average, were chosen and 624 sentences were successfully parsed with an average of 13.5 ambiguities.

D. Resultant Stochastic Context-free Grammar

After a certain number of iterations, probabilities were successfully associated to all of the grammar rules and the lexical rules as shown below:

```
* IT4
          0.98788          HELP     ---(a)
          0.00931          SEE      ---(b)
          0.00141          HEAR
          0.00139          WATCH
          0.00000          HAVE
          0.00000          FEEL
                    .
                    .
                    .
* SE
          0.28754          PRN VX PD      ---(c)
          0.25530          AAA 4X VX PD ---(d)
          0.14856          NNN VX PD
          0.13567          AV1 SE
          0.04006          PRE NQ SE
          0.02693          AV4 IX MX PD
          0.01714          NUM 4X VX PD
          0.01319          IT1 N2 PD
                    .
                    .
                    .
* VX
          0.16295          VT1 N2
          0.14372          VI1
          0.11963          AUX BV
          0.10174          PRE NQ VX
          0.09460          BE3 PA
                    .
```

In the above list, (a) means that "HELP" will be generated from part-of-speech "IT4" with the probability 0.98788, and (b) means that "SEE" will be generated from part-of-speech "IT4" with the probability 0.00931. (c) means that the non-terminal "SE (sentence)" will generate the sequence, "PRN (pronoun)", "VX (predicate)" and "PD (period or post sentential modifiers followed by period)" with the probability 0.28754. (d) means that "SE" will generate the sequence, "AAA (article, adjective, etc.)" , "4X (subject noun phrase)", "VX" and "PD" with the probability 0.25530. The remaining lines are to be interpreted similarly.

E. Parse Trees with Probabilities

Parse trees were printed as shown below including relative probabilities of each parse.

```
WE DO NOT UTILIZE OUTSIDE ART SERVICES DIRECTLY .

** total ambiguity is :      3

 *:       SENTENCE
   *:       PRONOUN       'we'
   *:       PREDICATE
     *:       AUXILIARY    'do'
     *:       INFINITE VERB PHRASE
       *        ADVERB TYPE1 'not'
       A: 0.356 INFINITE VERB PHRASE
        |*:       VERB TYPE IT1'utilize'
        |*:       OBJECT
        | *:        NOUN           'outside'
        | *:        ADJ CLAUSE
        |   *:        NOUN           'art'
        |   *:        PRED. WITH NO OBJECT
        |     *:        VERB TYPE VT1 'services'
       B: 0.003 INFINITE VERB PHRASE
        |*:       VERB TYPE IT1'utilize'
        |*:       OBJECT
        | *:        PREPOSITION  'outside'
        | *:        NOUN OBJECT
        |   *:        NOUN           'art'
        | *:        OBJECT
        |   *:        NOUN           'services'
       C: 0.641 INFINITE VERB PHRASE
        |*:       VERB TYPE IT1'utilize'
        |*:       OBJECT
        | *:        NOUN           'outside'
        | *:        OBJECT MASTER
        |   *:        NOUN           'art'
        |   *:        OBJECT MASTER
        |     *         NOUN           'services'
   *:       PERIOD
     *:       ADVERB TYPE1 'directly'
     *:       PRD           '.'
```

This example shows that the sentence 'We do not utilize outside art services directly.' was parsed in three different ways. The differences are shown as the difference of the sub-trees identified by A, B and C in the figure.

The numbers following the identifiers are the relative probabilities. As shown in this case, the correct parse, the third one, got the highest relative probability, as was expected.

F. Result

63 ambiguous sentences from magazine corpus and 21 ambiguous sentences from IBM correspondence were chosen at random from the sample sentences and their parse trees with probabilities were manually examined as shown in the table below:

a.	Corpus	Magazine	IBM
b.	Number of sentences checked manually	63	21
c.	Number of sentences with no correct parse	4	2
d.	Number of sentences which got highest prob. on most natural parse	54	18
e.	Number of sentences which did not get the highest prob. on the most natural parse	5	1
f.	Success ratio d/(d+e)	.915	.947

Taking into consideration that the grammar is not tailored for this experiment in any way, the result is quite satisfactory.

The only erroneous case of the IBM corpus is due to a grammar problem. That is, in this grammar, such modifier phrases as TO-infinitives, prepositional phrases, adverbials, etc. after the main verb will be derived from the 'end marker' of the sentence, i.e. period, rather than from the relevant constituent being modified. The parse tree in the previous figure is a typical example, that is, the adverb 'DIRECTLY' is derived from the 'PERIOD' rather than from the verb 'UTILIZE'. This simplified handling of dependencies will not keep information between modifying and modified phrases and as a result, will cause problems where the dependencies have crucial roles in the analysis. This error occurred in a sentence ' **... is going to work out**', where the two interpretations for the phrase '**to work**' exist:

'**to work**' modifies 'period' as:

1. A TO-infinitive phrase
2. A prepositional phrase

Ignoring the relationship to the previous context '**is going**', the second interpretation got the higher probability because prepositional phrases occur more frequently than TO-infinitive phrases if the context is not taken into account.

IV. CONCLUSION

The result from the trials suggests the strong potential of this method. And this also suggests some application possibility of this method such as: refining, minimizing, and optimizing a given context-free grammar. It will be also useful for giving a disambiguation capability to a given ambiguous context-free grammar.

In this experiment, an existing grammar was picked with few modifications, therefore, only statistics due to the syntactic differences of the sub-structured units were gathered. Applying this method to the collection of statistics which relate more to semantics should be investigated as the next step of this project. Introduction into the grammar of a dependency relationship among sub-structured units, semantically categorized parts-of-speech, head word inheritance among sub-structured units, etc. might be essential for this purpose. More investigation should be done on this direction.

V. ACKNOWLEDGEMENTS

This work was carried out when the author was in the Computer Science Department of the IBM Thomas J. Watson Research Center.

The author would like to thank Dr. John Cocke, Dr. F. Jelinek, Dr. B. Mercer, Dr. L. Bahl of the IBM Thomas J. Watson Research Center, and Prof. S. Kuno, of Harvard University for their encouragement and valuable technical suggestions.
Also the author is indebted to Mr. E. Black, Mr. B. Green and Mr. J Lutz for their assistance and discussions.

VII. REFERENCES

- Bahl,L.,Jelinek,F., and Mercer,R.,A Maximum Likelihood Approarch to Continuous Speech Recognition,Vol. PAMI-5,No.2, IEEE Trans. Pattern Analysis and Machine Intelligence,1983
- Bahl,L.,et.al.,Automatic Recognition of Continuously Spoken Sentences from a finite state grammar, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Tulsa, OK, Apr. 1978
- Bahl,L.,et.al.,Further results on the recognition of a continuousl read natural corpus, Proc. IEEE Int. Conf. Acoust., Speech,Signal Processing,Denver,CO,Apr. 1980
- Baum,L.E.,A Maximazation Technique occurring in the Statistical Analysis of Probablistic Functions of Markov Chains, Vol. 41, No.1, The Annals of Mathematical Statistics, 1970
- Baum,L.E.,An Inequality and Associated Maximization Technique in Statistical Estimation for Probablistic Functions of Markov Processes, Inequalities, Vol. 3, Academic Press, 1972
- Fu,K.S.,Syntactic Methods in Pattern Recognition,Vol 112, Mathematics in science and Engineering, Academic Press, 1974
- Knuth,D.,Fundamental Algorithms,Vol 1. in The Art of Computer Programming, Addison Wesley, 1975
- Kuno,S.,The Augmented Predictive Analyzer for Context-free Languages-Its Relative Efficiency, Vol. 9, No. 11, CACM, 1966
- Kuno,S.,Oettinger,A.G.,Syntactic Structure and Ambiguity of English, Proc. FJCC, AFIPS, 1963
- Martin,W., et.al.,Preliminary Analysis of a Breadth-First Parsing Algorithm:Theoretical and Experimental Results, MIT LCS report TR-261, MIT 1981
- Sibuya,M.,Fujisaki,T. and Takao,Y.,Noun-Phrase Model and Natural Query Language, Vol 22, No 5,IBM J. Res. Dev. 1978
- Winograd,T.,Understanding Natural Language, Academic Press, 1972
- Woods,W.,The Lunar Sciences Natural Language Information System, BBN Report No. 2378, Bolt, Beranek and Newman