A PROBABILISTIC PARSER

Roger Garside and Fanny Leech
Unit for Computer Research on the English Language
University of Lancaster
Bailrigg
Lancaster LA1 4YT, U.K.

ABSTRACT

The UCREL team at the University of Lancaster is engaged in the development of a robust parsing mechanism, which will assign the appropriate grammatical structure to sentences in unconstrained English text. The techniques used involve the calculation of probabilities for competing structures, and are based on the techniques successfully used in tagging (i.e. assigning grammatical word classes) to the LOB (Lancaster-Oslo/Bergen) corpus.

The first step in the parsing process involves dictionary lookup of successive pairs of grammatically tagged words, to give a number of possible continuations to the current parse. Since this lookup will often not be able unambiguously to distinguish the point at which a grammatical constituent should be closed, the second step of the parsing process will have to insert closures and distinguish between alternative parses. It will generate trees representing these possible alternatives, insert closure points for the constituents, and compute a probability for each parse tree from the probability of each constituent within the tree. It will then be able to select a preferred parse or parses for output.

The probability of a grammatical constituent is derived from a bank of manually parsed sentences.

INTRODUCTION

In this paper we present an overview of one part of the work currently being carried out at the Unit for Computer Research on the English Language (UCREL) in the University of Lancaster, under SERC research grant number GR/C/47700. This work involves the automatic syntactic analysis or parsing of the LOB corpus, using the statistical or constituent-likelihood (CL) grammar ideas of Atwell (1983). The work is based on the grammatical tagging of the LOB corpus, both as providing a partially analysed text and because of the techniques used in assigning tags. We therefore begin by briefly describing this earlier project.

The grammatical tagging of the LOB corpus is described in detail elsewhere (see, for example, Leech, Garside and Atwell 1983, Marshall 1983, Beale 1985), but in essence there are three stages. The first stage takes the original corpus, on which a certain amount of pre-editing (both automatic and manual) has been performed. It assigns to each word in the corpus a set of possible tags, and it is assumed that the correct tag is in this set. The set of possible tags is chosen without at this stage considering the context in which the word appears, and the choice is made by using an ordered set of decision rules, the most commonly used of which (in about 65-70% of cases) is to look the word up in a dictionary of some 7000 words.

The third stage involves looking at those cases where the first stage has resulted in more than one tag being assigned to a word. In this case we calculate the probability of each possible sequence of ambiguous tags, and the most likely sequence is chosen as the correct one. In most cases the probability of a sequence of tags is calculated by multiplying together the pairwise probabilities of one tag following another, and these pairwise probabilities were derived from a statistical analysis of co-occurrence of tags in the tagged Brown corpus (Francis and Kucera 1964).

A further stage was later inserted between the two stages described above. This stage involves the ability to look for patterns of sequences of words and putative tags assigned by the first stage, and to modify the sets of tags assigned to words. This enables various problematical situations to be resolved or clarified in order to improve the disambiguating ability of the third stage.

After the third stage (when the appropriate tag will have been automatically selected some 96.5% of the time), the remaining errors are removed by a manual post-editing phase.

The fundamental idea on which our syntactic analysis is based, originally formulated in Atwell (1983), is that the general principles behind the tagging system could be used at the parsing level. Thus a first stage of parsing could be to look up a tag in a dictionary to derive a set of possible constituents (or "hypertags") containing this tag. Similarly, in the third stage, the probability of any particular constituent being constructed out of a particular set of constituents or word-classes at the next lower level could be used to disambiguate a set of constituents posited at the first stage. To this end some 2000 sentences from the LOB corpus have been manually parsed, and the results stored as a "treebank" or database of information on the frequency of occurrence of possible grammatical structures. Thus, for each possible "mother" constituent, there will be stored a set of sequences of daughter constituents or word-classes, together with their frequencies.

The second stage generalises to a search for particular syntactic patterns which are recognisable in context, and the resolution of which will improve the accuracy of the third stage. We develop these ideas in the remainder of the paper.

INPUT TO THE ANALYSIS SYSTEM

The input to the analysis system is essentially the output from the tagging system described above. An example of this is given in figure 1.

```
BO1   9 001 ------------------------------------
BO1   9 010 there                          EX
BO1   9 020 is                             BEZ
BO1   9 030 the                            AT
BO1   9 040 possibility                    NN
BO1   9 050 that                           CS
BO1   9 060 it                             PP3
BO1   9 070 will                           MD
BO1   9 080 not                            XNOT
BO1   9 090 be                             BE
BO1   9 100 settled                        VBN
BO1   9 110 at                             IN
BO1   9 120 this                           DT
BO1  10 010 conference                     NN
BO1  10 011 .                              .
BO1  10 012 ------------------------------------
```

Figure 1. Input to the System.

Each line of the tagged LOB corpus contains one word or punctuation mark, and each sentence is separated from the preceding one by the sentence initial marker, here represented by a horizontal line. Each line consists of three main fields; a reference number specifying the genre, text number, line number, and position within the line; the word or punctuation mark itself; and the correct tag. The tags are taken from a set of 134 tags, based on the Brown tagset (Greene and Rubin 1971), but modified where we felt it was desirable.

OUTPUT FROM THE ANALYSIS SYSTEM

Typical output from the analysis system would look like figure 2.

```
BO1   9 001 ------------------------------------
BO1   9 010 there               EX      [S[E]
BO1   9 020 is                  BEZ     [V]
BO1   9 030 the                 AT      [N
BO1   9 040 possibility         NN
BO1   9 050 that                CS      [Fn
BO1   9 060 it                  PP3     [N]
BO1   9 070 will                MD      [Ve
BO1   9 080 not                 XNOT
BO1   9 090 be                  BE
BO1   9 100 settled             VBN     Ve]
BO1   9 110 at                  IN      [P
BO1   9 120 this                DT      [N
BO1  10 010 conference          NN      N]P]Fn]N]
BO1  10 011 .                   .       S]
BO1  10 012 ------------------------------------
```

Figure 2. Output from the System.

The field on the right is meant to represent a typical parse tree, but in a columnar form. Each constituent is represented by a an upper case letter; thus S is the sentence, N is a noun phrase, and F indicates a subordinate clause. The upper case letter may be followed by one or more lower case letters, indicating features of interest in the constituent; thus Fn indicates a nominal clause. The boundaries of a constituent are given by open and close square brackets, so that for instance the subordinate clause indicated by Fn starts at the word "that" and ends at the word "conference".

STAGE ONE - ASSIGNMENT

It is clear that a tag, or a pair of consecutive tags, is partially diagnostic of the beginning, continuation or termination of a constituent. Thus, for example, the pair "noun-verb" tends to indicate the end of a noun phase and the beginning of a verb phase, and the pair "noun-noun" tends to indicate the continuation of a noun phase. The first step in the syntactic analysis is therefore to deduce from the sequence of tags a tentative sequence of markings for the type and boundaries of the constituents. Since the beginnings of constituents tend to be marked, but not the ends, this sequence of markings will tend to omit many of the right-hand or closing brackets, and these are inserted at a later stage.

The first stage of parsing is therefore to look up each (tag, tag) pair in a dictionary, and this results in one or more possible sequences of open and close brackets and constituent markings - each of these sequences is, for historical reasons, called a "T-tag". A T-tag consists of a left-hand and a right-hand part. The left-hand part consists of an indication of what constituent should be current (i.e. at the top of the stack of open constituents) at this stage, perhaps followed by one or more closing brackets. The right-hand part normally consists of an indication that one or more new constituents should be opened, that some particular constituent should be continued, or more rarely that a new constituent should be (and this will be deduced later on in the analysis process). Thus the tag-pair "noun followed by subordinating conjunction" indicates two possible T-tags, either "Y] [F" or "Y [F". The first means close the current constituent whatever it is (Y matches any constituent) and open a new subordinate clause (F) constituent, while the second means continue the current constituent and open an F constituent.

The look-up procedure as described above requires a dictionary entry for each possible pair of tags, which is inefficient and difficult to relate to meaningful linguistic categories. Instead the 134 tags are subsumed in a set of 33 "cover symbols" (the term is taken from the Brown tagging system). Thus all the different forms of noun word tag are subsumed in the cover symbols N* (singular noun), *S (plural noun) and *$ (noun with genitive marker). The required tag-pair dictionary will therefore require only an entry for each cover-symbol pair (together with a list of exceptions, where the tag rather than the cover symbol is diagnostic of the appropriate T-tags). A further simplification is that in many cases (because of the admissibility of the "wild" constituent marker Y) the first tag of the pair is irrelevant and the second tag in the pair determines the set of T-tag options.

I said that the T-tag dictionary look-up would often result in more than one possible T-tag, rather than just one. Some of these options can be eliminated immediately by matching the current constituent with the putative extension, but others need to be retained for later disambiguation.

CONSTRUCTING THE T-TAG DICTIONARY

The original version of the T-tag dictionary was generated using linguistic intuition. If there are several possible T-tags to an entry, they are given in approximately decreasing likelihood and rare T-tags are marked as such. The treebank of manually parsed sentences can now be used to extract information about what constituent types and boundaries are associated with what pairs of tags. We have therefore written a program which takes a current version of the T-tag dictionary and a set of parsed sentences, and generates;

(a) information about putative exceptions to the curent T-tag dictionary, in the form of cases where the effective T-tag in the parsed sentence is not among those proposed by the T-tag dictionary, and

(b) where the effective T-tag _is_ among those proposed by the T-tag dictionary, statistics are gathered as to the differential probabilities of the various T-tags associated with a particular tagpair.

The first set of information is used to guide the intuition of a linguist in deciding how to modify the original T-tag table. This cannot (at least at present) be done automatically, since there are various unsystematic differences between the T-tag as looked up in the dictionary and the sequence of constituent types and boundaries as they appear in the parsed sentences. We are thus using information from the parsed corpus texts to generate improved versions of the T-tag dictionary.

The frequency information about the optional T-tags associated with a particular tagpair is not at present used by the analysis system, but we feel that it may be a further factor to be taken into account when deciding on a preferred parse in the third stage of analysis. The information is of course being used to refine linguistic intuition about the ordering of possible T-tags in the dictionary and their marking for rarity.

STAGE THREE - TREE-CLOSING

The output from the first stage consists of indications of a number of constituents and where they begin, but in many cases the ending position of a constituent is unknown, or at least is located ambiguously at one of several positions. The main task of the third stage is to insert these constituent closures. There is a further stage between T-tag assignment and tree-closure which we will return to in a later section.

The third stage proceeds as follows. A backward search is made from the end of the sentence to find a position at which choices and/or decisions have to be made. At the first such point the alternative trees are constructed and then all unclosed constituents are completed, by means of likelihood calculations based on the database of probabilities. To effect closure, the last unclosed constituent is selected and a subtree data structure is created to represent this constituent. The parser then attempts to attach to it as daughters any constituents (word-classes or constituents) lying positionally below it. As a consequence of each successive attachment there exists a distinct mother-daughter sequence pattern, the probability of which can be extracted from the mother-daughter table derived from the treebank (the parser will not attempt to build subtrees with probabilities below a certain threshold). If a sequence of constituents is attached as daughters, then any remaining constituents lying below the last attached daughter are attached to the subtree as sisters. Thus the constituent is closed in all statistically possible ways, and the parser is once again positioned at the end of the sentence.

The parser again selects the next unclosed constituent, this time passing over the newly closed constituent (which is now represented as a subtree), and it proceeds to close the new constituent in the manner described above. However when attaching as daughter or sister the newly closed constituent from the previous selection it attaches a set of subtrees that represents all its possible closure patterns. This process is repeated until the top level is reached. If the head of the sentence has been reached, then many sub-trees are discarded because at this level all other constituents must be daughters and not sisters. If more than one tree is to be completed from a choice, then this process is repeated until all the alternative trees have been closed.

STATISTICS FOR THE MOTHER-DAUGHTER SEQUENCES

The main problem is how to store the frequency information on possible daughter sequences for each mother constituent. Originally the manually parsed sentences collected in the treebank were decomposed into a mother constituent and each of its daughter sequences in its entirety. So for a mother constituent N (noun phrase) a possible daughter is "ATI, JJ, NNS, Fr" (i.e. determiner, adjective, plural noun, subordinate clause).

The main problem with this is that, for all the most common daughter sequences, the statistics were too dependent on exactly which sentences had occurred. This also implies that the parser has to match very specific patterns when a subtree is being investigated. To produce statistical tables of sufficient generality, each daughter sequence was decomposed into its individual pairs of elements (each daughtser sequence in its entirety having implied opening and closing delimiters, represented by the symbols '[' and ']' respectively) and all like pairs were added together. The frequency information now consists of the mother constituent and a set of daughter pairs.

Now, for the parser to assess the probability of any daughter sequence, this sequence has first to be decomposed into pairs, which are looked up in the mother-daughter table, and the probabilities of the pairs aggregated together to give the overall probability of the complete sequence. For the sequence described above the individual pairs would be "[ATI, ATI JJ, JJ NNS, NNS Fr, Fr]".

It seems clear that in some cases the aggregation of the probabilities of two or more pairs does not give a reasonable approximation to the original statistics, because of longer-distance dependencies. It is likely therefore that this technique will need a dictionary of pairs together with a dictionary of exceptional triples, quadruples, etc., to correct the pairs dictionary where necessary.

STAGE TWO - HEURISTICS

The first stage of T-tag assignment introduces constituent types and boundary markings only if they can be expressed in terms of look-up in a dictionary of tag-pairs. However there are a number of cases where a more complex form of processing seems desirable, in order to produce a more suitable partial parse to be fed to the third stage. We are therefore designing a second stage, analogous to the second stage of the tagging system, which is able to look for various patterns of tags and the constituent markings already assigned by the first stage, and then add to or modify the constituent markings passed to the third stage; an area where this will be important is in coordinated structures.

I have suggested in the above that the parsing system is constructed as three separate stages, which pass their output to the next stage. In fact this is mainly for expository and developmental reasons, and we envisage an interconnection between at least some of the stages, so that earlier stages may be able to take account of information provided by later stages.

PROBLEMS AND CONCLUSIONS

I have described the basic structure of the parsing system that we are currently developing at Lancaster. There are of course a number of areas where the techniques described will need to be extended to take account of lingustic structures not provided for. But our technique with the tagging project was to develop basic mechanisms to cope with a large portion of the texts being processed, and then to modify then to perform more accurately in particular areas where they were deficient, and we expect to follow this procedure with the current project.

The two main features of the technique we are using seem to be

(a) the use of probabilistic methods for disambiguation of linguistic structures, and

(b) the use of a corpus of unconstrained English text as a testbed for our methods, as a source of information about the statistical properties of language, and as an indicator of what are the important areas of inadequacy in each stage of the analysis system.

Because of the success of these techniques in the tagging system, and because of the promising results already achieved in applying these techniques to the syntactic analysis of a number of simple sentences, we have every hope of being able to develop a robust and economic parsing system able to operate over unconstrained English text with a high degree of accuracy.

REFERENCES

Atwell, E.S. (1983), "Constituent-Likelihood Grammar". Newsletter of the International Computer Archive of Modern English (ICAME News) 7, 34-66.

Beale, A.D. (1985), "Grammatical Analysis by Computer of the Lancaster-Oslo/Bergen (LOB) Corpus of British English Texts". Proceedings of the Second ACL European Conference (To appear).

Francis, W.N. and Kucera, H. (1964), "Manual of Information to Accompany a Standard Sample of Present-Day Edited American English, for Use with Digital Computers". Department of Linguistics, Brown University.

Greene, B.B. and Rubin, G.M. (1971). "Automatic Grammatical Tagging of English". Department of Linguistics, Brown University.

Leech, G.N., Garside, R.G. and Atwell, E.S. (1983). "The Automatic Grammatical Tagging of the LOB Corpus". Newsletter of the International Computer Archive of Modern English (ICAME News) 7, 13-33.

Marshall, I. (1983), "Choice of Grammatical Word-Class without Global Syntactic Analysis: Tagging Words in the LOB Corpus". Computers and the Humanities 17, 139-50.