Parsing with Principles and Probabilities

Andrew Fordham
SCS Research Group
Department of Sociology
University of Surrey
Guildford
Surrey, GU2 5XH, UK
`ajf@soc.surrey.ac.uk`

Matthew Crocker
Centre for Cognitive Science
University of Edinburgh
2 Buccleuch Place
Edinburgh, EH8 9LW
Scotland
`mwc@cogsci.ed.ac.uk`

Abstract

This paper is an attempt to bring together two approaches to language analysis. The possible use of probabilistic information in principle-based grammars and parsers is considered, including discussion on some theoretical and computational problems that arise. Finally a partial implementation of these ideas is presented, along with some preliminary results from testing on a small set of sentences.

Introduction

Both principle-based parsing and probabilistic methods for the analysis of natural language have become popular in the last decade. While the former borrows from advanced linguistic specifications of syntax, the latter has been more concerned with extracting distributional regularities from language to aid the implementation of NLP systems and the analysis of corpora.

These symbolic and statistical approaches are beginning to draw together as it becomes clear that one cannot exist entirely without the other: the knowledge of language posited over the years by theoretical linguists has been useful in constraining and guiding statistical approaches, and the corpora now available to linguists have resurrected the desire to account for real language data in a more principled way than had previously been attempted.

This paper falls directly between these approaches, using statistical information derived from corpora analysis to weight syntactic analyses produced by a 'principles and parameters' parser. The use of probabilistic information in principle-based grammars and parsers is considered, including discussion on some theoretical and computational problems that arise. Finally a partial implementation of these ideas is presented, along with some preliminary results from testing on a small set of sentences.

Government-Binding Theory

The principles and parameters paradigm in linguistics is most fully realised in the Government-Binding Theory (GB) of Chomsky [Chomsky1981, Chomsky1986] and others. The grammar is divided into modules which filter out ungrammatical structures at the various levels of representation; these levels are related by general transformations. A sketch of the organisation of GB (the 'T-model') is shown in figure 1.

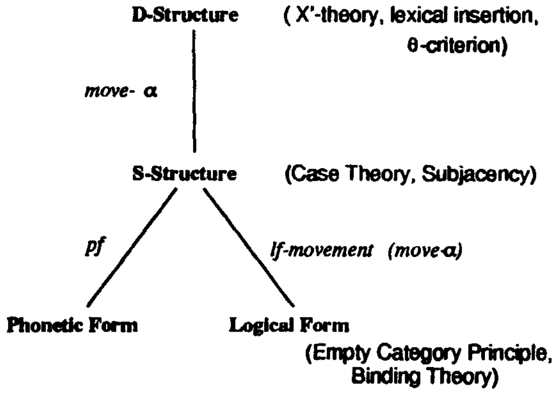

Figure 1: The T-model of grammar

Little work has been done on the complexity of algorithms used to parse with a principle-based grammar, since such grammars do not exist as accepted mathematically well-defined constructs. It has been estimated that in general, principle-based parsing can only be accomplished in exponential time, *i.e.* $O(2^n)$ [Berwick and Weinberg1984, Weinberg1988].

A feature of principle-based grammars is their potential to assign some meaningful representation to a string which is strictly ungrammatical. It is an inherent feature of phrase structure grammars that they classify the strings of words from a language into two (infinite) sets, one containing the grammatical strings and the other the ungrammatical strings. Although attempts have been made to modify PS grammars/parsers to cope with extragrammatical input, *e.g.* [Carbonell and Hayes1983, Douglas and Dale1992, Jensen et al.1983, Mellish1989], this is a feature which has to be 'added on' and tends to affect the statement of the grammar.

Due to the lack of an accepted formalism for the specification of principle-based grammars, Crocker and Lewin [Crocker and Lewin1992] define the declarative 'Proper Branch' formalism, which can be used with a number of different parsing methods.

A proper branch is a set of three nodes — a mother and two daughters — which are constructed by the parser, using a simple mechanism such as a shift-reduce interpreter, and then 'licensed' by the principles of grammar. A complete phrase marker of the input string can then be constructed by following the manner in which the mother node from one proper branch is used as a daughter node in a dominating proper branch.

Each proper branch is a binary branching structure, and so all grammatical constraints will need to be encoded locally. Crocker [Crocker1992] develops "a 'representational' reformulation of the transformational model which decomposes syntactic analysis into several representation types — including phrase structure, chains, and coindexation — allowing one to maintain the strictly local characterisation of principles with respect to their relevant representation types," [Crocker and Lewin1992, p. 511].

By using the proper branch method of axiomatising the grammar, the structure building section of the parser is only constrained in that it must produce proper branches; it is therefore possible to experiment with different interpreters (*i.e.* structure proposing engines) while keeping the grammar constant.

The Grammar and Parser

A small principle-based parser was built, following the proper branch formalism developed in [Crocker and Lewin1992]. Although the grammar is very limited, the use of probabilities in ranking the parser's output can be seen as a first step towards implementing a principle-based parser using a more fully specified collection of grammar modules.

The grammar is loosely based on three modules taken from Government-Binding Theory — X-bar theory, Theta Theory and Case Theory. Although these embody the spirit of the constraints found in Chomsky [Chomsky1981] they are not intended to be entirely faithful to this specification of syntactic theory. There is also only a single level of representation (which is explicitly constructed for output purposes but not consulted by the parser). This representation is interpreted as S-structure.

Explanations of the knowledge contained within each grammar principle is given in the following sections.

$\bar{X}$ Theory

X-bar Theory uses a set of schemata to license local subtrees. We use a parametrised version of the X-bar schemata, similar to that of Muysken [Muysken1983], but employing features which relate to the state of the head word's theta grid to give five schemata (figure 2). A node includes the following features (among others):

1. $X^{-}_{-} \rightarrow Y^{-}_{-}\ X^{+}_{-}$
2. $X^{S}_{+} \rightarrow X^{S}_{+}\ Y^{+}_{-}$
3. $X^{S}_{-} \rightarrow X^{S}_{+}\ Y^{-}_{-}$
4. $X^{S}_{-} \rightarrow X^{S}_{-}\ Y^{-}_{-}$
5. $X^{S}_{-} \rightarrow Y^{-}_{-}\ X^{S}_{-}$

Figure 2: The X-bar Schemata

1. Category: the standard category names are employed.
2. Specifier (SPEC): this feature specifies whether the word at the head of the phrase being built requires a specifier.
3. Complement (COMP): the complement feature is redundant in that the information used to derive it's value is already present in a word's theta grid, and will therefore be checked for well-formedness by the theta criterion. Since this information is not referenced until later, the COMP feature is used to limit the number of superfluous proper-branches generated by the parser.
4. The head (*i.e.* lexical item) of a node is carried on each projection of that node along with its theta grid.

The probabilities for occurrences of the X-bar schema were obtained from sentences from the preliminary Penn Treebank corpus of the Wall Street Journal, chosen because of their length and the head of their verb phrase (*i.e.* the main verbs were all from the set for which theta role data was obtained); the examples were manually parsed by the authors.

The probabilities were calculated using the following equation, where $X^{S_i}_{C_i} \rightarrow Y^{S_j}_{C_j}\ Z^{S_k}_{C_k}$ is a specific schema, $\bar{X}$ is the set of X-bar schemata and A and B and C are variables over category, SPEC and COMP feature bundles:

$$P(X^{S_i}_{C_i} \rightarrow Y^{S_j}_{C_j}\ Z^{S_k}_{C_k} | \bar{X}) = \frac{C(X^{S_i}_{C_i} \rightarrow Y^{S_j}_{C_j}\ Z^{S_k}_{C_k})}{C(A \rightarrow B\ C)} \quad (1)$$

This is different to manner in which probabilities are collected for stochastic context-free grammars, where the identity of the mother node is taken into account, as in the equation below:

$$P(X^{S_i}_{C_i} \rightarrow Y^{S_j}_{C_j}\ Z^{S_k}_{C_k} | \bar{X}) = \frac{C(X^{S_i}_{C_i} \rightarrow Y^{S_j}_{C_j}\ Z^{S_k}_{C_k})}{C(X^{S_i}_{C_i} \rightarrow B\ C)} \quad (2)$$

This would result in misleading probabilities for the X-bar schemata since the use of schemata (3), (4), and (5) would immediately bring down the probability of a parse compared to a parse of the same string which happened to use only (1) and (2).[1]

---

[1] The probabilities for (1) and (2) would be 1 as they have unique mothers.

The overall (X-bar) likelihood of a parse can then be computed by multiplying together all the probabilities obtained from each application of the schemata, in a manner analogous to that used to obtain the probability of a phrase marker generated by an SCFG. Using the schemata in this way suggests that the building of structure is category independent, *i.e.* it is just as likely that a verb will have a (filled) specifier position as it is for a noun. The work on stochastic context-free grammars suggests a different set of results, in that the specific categories involved in expansions are all important. While SCFGs will tend to deny that all categories expand in certain ways with the same probabilities, they make this claim while using a homogeneous grammar formalism. When a more modular theory is employed, the source of the supposedly category specific information is not as obvious. The use of lexical probabilities on specifier and complement co-occurrence with specific heads (*i.e.* lexical items) could exihibit properties that appear to be category specific, but are in fact caused by common properties which are shared by lexical items of the same category.[2] Since it can be argued that the probabilistic information on lexical items will be needed independently, there is no need to use category specific information in assigning probabilities to syntactic configurations.

Theta Theory

Theta theory is concerned with the assignment of an argument structure to a sentence. A verb has a number of the thematic (or 'theta') roles which must be assigned to its arguments, *e.g.* a transitive verb has one theta role to 'discharge' which must be assigned to an NP.

If a binary branching formalism is employed, or indeed any formalism where the arguments of an item and the item itself are not necessarily all sisters, the problem of when to access the probability of a theta application is presented. The easiest method of obtaining and applying theta probabilities will be with reference to whole theta grids. Each theta grid for a word will be assigned a probability which is not dependent on any particular items in the grid, but rather on the occurrence of the theta grid as a whole.

A preliminary version of the Penn Treebank bracketed corpus was analysed to extract information on the sisters of particular verbs. Although the Penn Treebank data is unreliable since it does not always distinguish complements from adjuncts, it was the only suitable parsed corpus to which the authors had access. Although the distinction between complements and adjuncts is a theoretically interesting one, the process of determining which constructions fill which functional roles in the analysis of real text often creates a number of problems (see [Hindle and Rooth1993] for discussion on this issue regarding output of the Fidditch parser [Hindle1993]).

The probabilities for each of the verbs' theta grids were calculated using the equation below, where $P(s_i|v)$ is the probability of the theta grid s_i occurring with the verb v, (v, s_i) is an occurrence of the items in s_i being licensed by v, and S ranges over all theta grids for v:

$$P(s_i|v) = \frac{C(v, s_i)}{C(v, S)} \tag{3}$$

Case Theory

In its simplest form, Case theory invokes the Case filter to ensure that all noun phrases in a parse are assigned (abstract) case. Case theory differs from both X-bar and Theta theory in that it is category specific: only NPs require, or indeed can be assigned, abstract case. If we are to implement a probabilistic version of a modular grammar theory incorporating a Case component, a relevant question is: are there multiple ways of assigning Case to noun phrases in a sentence? *i.e.* can ambiguity arise due to the presence of two candidate Case assigners?

Case theory suggests that the answer to this is negative, since Case assignment is linked to theta theory *via* visibility, and it is not possible for an NP to receive more than one theta role. As a result, the use of Case probabilities in a parser would be at best unimportant, since some form of ambiguity is needed in the module, *i.e.* it is possible to satisfy the Case filter in more than one way, for probabilities associated with the module to be of any use. While having a provision for using probabilities deduced from Case information, the implemented parser does not in fact use Case in its parse ranking operations.

Local Calculation

The use of a heterogeneous grammar formalism and multiple probabilities invokes the problem of their combination. There are at least two ways in which each mother's probabilities can be calculated; firstly, the probability information of the same type can be used: the daughters' X-bar probabilities alone could be used in calculating the mother's X-bar probability. Alternatively, a combination of some or all of the daughters' probability features could be employed, thus making, *e.g.*, the X-bar probability of the mother dependent upon all the stochastic information from the daughters, including theta and Case probabilities, *etc.*

The need for a method of combining the daughter probabilities into a useful figure for the calculation of the mother probabilities is likely to involve trial and error, since theory thus far has had nothing to say on the subject. The former method, using only the relevant daughter probabilities, therefore seems to be the most fruitful path to follow at the outset, since it does not require a way of integrating probabilities from different modules while the parse is in progress, nor is it as computationally expensive.

---

[2] It is of course possible to store these cross-item similarities as lexical rules [Bresnan1978], but this alone does not entail that the properties are specific to a category, *cf.* the theta grids of verbs and their 'related' nouns.

Global Calculation

The manner in which the global probability is calculated will be partly dependent upon the information contained in the local probability calculations.

If the probabilities for partial analyses have been calculated using only probabilities of the same types from the subanalyses — *e.g.* X-bar, Theta — the probabilities at the top level will have been calculated using informationally distinct figures. This has the advantage of making 'pure' probabilities available, in that the X-bar probability will reflect the likelihood of the structure alone, and will be 'uncontaminated' by any other information. It should then be possible to experiment with different methods of combining these probabilities, other than the obvious 'multiplying them together' techniques, which could result in one type of probabililty emerging as the most important.

On the other hand, if probabilities calculated during the parse take all the different types of probabilities into account at each calculation — *i.e.* the X-bar, theta, *etc.* probabilities on daughters are all taken into account when calculating the mother's X-bar probability — the probabilities at the top level will not be pure, and a lot of the information contained in them will be redundant since they will share a large subset of the probabilities used in their separate calculations. It will not therefore be easy to gain theoretical insight using these statistics, and their most profitable method of combination is likely to be more haphazard affair than when more pure probabilities are used.

The parser used in testing employed the first method and therefore produced separate module probabilities for each node. For the lack of a better, theoretically motivated method for combining these figures, the product of the probabilities was taken as the global probability for each parse.

Testing the Parser

The parser was tested using sixteen sentences containing verbs for which data had been collected from the Penn Treebank corpus. The sentences were created by the authors to exhibit at least a degree of ambiguity when it came to attaching a post-verbal phrase as an adjunct or a complement. In order to force the choice of the 'best' parse on to the verb, the probabilities of theta grids for nouns, prepositions, *etc.* was kept constant.

Of these 16 highest ranked parses, 7 are the expected parse, with the other 9 exhibiting some form of misattachment. The fact that each string received multiple parses (the mean number of analyses being 9.135, and the median, 6) suggests that the probabilistic information did favourably guide the selection of a single analysis.

It is not really possible to say from these results how successful the whole approach of probabilistic principle-based parsing would be if it were fully implemented. The inconclusive nature of the results obtained was due to a number of limiting factors of the implementation including the simplicity of the grammar and the lack of available data.

Discussion

Limitations of the Grammar

The grammar employed is a partial characterisation of Chomsky's Government-Binding theory [Chomsky1981, Chomsky1986] and only takes account of very local constraints (*i.e.* X-bar, Theta and Case); a way of encoding all constraints in the proper branch formalism (*e.g.* [Crocker1992]) will be needed before a grammar of sufficient coverage to be useful in corpora analysis can be formulated. The problem with using results obtained from the implementation given here is that the grammar is sufficiently underspecified and so leaves too great a task for the probabilistic information.

This approach could be viewed as putting the cart before the horse; the usefulness of stochastic information in parsers presumes that a certain level of accuracy can be achieved by the grammar alone. While GB is an elegant theory of cognitive syntax, it has yet to be shown that such a modular characteristion can be successfully employed in corpus analysis.

Statistical Data and their Source

The use of the preliminary Penn Treebank corpus for the extraction of probabilities used in the implementation above was a choice forced by lack of suitable materials. There are still very few parsed corpora available, and none that contain information which is specified to the level required by, *e.g.*, a GB grammar. While this is not an absolute limitation, in that it is theoretically possible to extract this information manually or semi-automatically from a corpus, time constraints entailed the rejection of this approach.

It would be ultimately desirable if the use of probabilities in principle-based parsing could be used to mirror the way that a syntactic theory such as Government-Binding handles constructions — various modules of the grammar conspire to rule out illegal structures or derivations. It would be an elegant result if a construction such as the passive were to use probabilities for chains, Case assignment *etc.* to select a parse that reflected the lexical changes that had been undergone, *e.g.* the greater likelihood of an NP featuring in the verb's theta grid. It is this property of a number of modules working hand in hand that needs to be carried over into the probabilistic domain.

The objections that linguists once held against statistical methods are disappearing slowly, partly due to results in corpora analysis that show the inadequacy of linguistic theory when applied to naturally occurring data. It is also the case that the rise of the connectionist phoenix has brought the idea of weighted (though not strictly probabilistic) functions of cognition back to the fore, freeing the hands of linguists who believe that while an explanatorily adequate theory of grammar is an elegant construct, its human implementation, and its usage in computational linguists may not be straight forward. This paper has hopefully shown that an integration of statistical methods and current linguistic theory is a goal worth pursuing.

References

[Berwick and Weinberg1984] Robert Cregar Berwick and Amy S. Weinberg. 1984. *The Grammatical Basis of Linguistic Performance: Language Use and Acquistion.* MIT Press.

[Bresnan1978] Joan. W. Bresnan. 1978. A realistic transformational grammar. In M. Halle, J. Bresnan, and G. Miller, editors, *Linguistic Theory and Psychological Reality.* MIT Press, Cambridge, MA.

[Carbonell and Hayes1983] Jaime G. Carbonell and Philip J. Hayes. 1983. Recovery strategies for parsing extragrammatical language. *American Journal of Computational Linguistics*, 9(3–4):123–146, July–December.

[Chomsky1981] Noam Chomsky. 1981. *Lectures on Government and Binding.* Studies in Generative Grammar No. 9. Foris, Dordrecht.

[Chomsky1986] Noam Chomsky. 1986. *Knowledge of Language: Its Nature, Origin, and Use.* Convergence. Praeger, New York.

[Crocker and Lewin1992] Matthew Walter Crocker and Ian Lewin. 1992. Parsing as deduction: Rules versus principles. In B. Neumann, editor, *ECAI 92. 10th European Conference on Artificial Intelligence*, pages 508–512. John Wiley and Sons, Ltd.

[Crocker1992] Matthew Walter Crocker. 1992. *A Logical Model of Competence and Performance in the Human Sentence Processor.* Ph.D. thesis, Dept. Artificial Intelligence, University of Edinburgh.

[Douglas and Dale1992] Shona Douglas and Robert Dale. 1992. Towards robust PATR. In Ch. Boitet, editor, *COLING-92, Proceedings of the fifteenth International Conference on Computational Linguistics*, pages 468–474.

[Hindle and Rooth1993] Donald Hindle and Mats Rooth. 1993. Structural ambuiguity and lexical relations. *Computational Linguistics*, 19(1).

[Hindle1993] Donald Hindle. 1993. A parser for text corpora. In B. T. S. Atkins and A. Zampolli, editors, *Computational Approaches to the Lexicon.*

[Jensen et al.1983] K. Jensen, G. E. Heidborn, L. A. Miller, and Y. Ravin. 1983. Parse fitting and prose fixing: Getting a hold on ill-formedness. *American Journal of Computational Linguistics*, 9(3–4):147–160, July–December.

[Mellish1989] Christopher S. Mellish. 1989. Some chart-based techniques for parsing ill-formed input. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 102-109.

[Muysken1983] Pieter Muysken. 1983. Parameterizing the notion of head. *Journal of Linguistic Research*, 2:57-76.

[Weinberg1988] Amy S. Weinberg. 1988. Mathematical properties of grammars. In Frederick J. Newmeyer, editor, *Linguistics: the Cambridge Survey, Vol. 1, Linguistics Theory: Foundations*, chapter 15, pages 415–429. Cambridge University Press.