ANTICIPATION-FREE DIAGNOSIS OF STRUCTURAL FAULTS

Wolfgang Menzel

Zentralinstitut für Sprachwissenschaft
Akademie der Wissenschaften der DDR
Prenzlauer Promenade 149-152
Berlin, 1100, DDR

Current attempts to diagnose grammatical faults in natural language utterances are - except for agreement errors and certain cases of overgeneralization and interference - strongly based on the principles of error anticipation (cf. Yazdani 1988, Schwind 1988, Catt 1988): Rather tiny context free grammars are enhanced by some additional rules which describe selected faulty structures and invoke error messages if they are needed for a successful parse. The efforts required to compile an at least approximatively comprehensive rule set even for simple domains of grammar are considerable. Besides this, it is the student's risk to fall into the remaining gap of neglected possibilities which seems to be difficult to avoid. Hopefully, an improvement of this situation can be achieved by an application of model-based reasoning procedures, where an internal model (of language correctness) is used to simulate and evaluate error hypotheses by investigating their consequences for other parts of the model. To a certain degree the diagnostic results are logically determined by the correct remainder of the utterance and useful results require a balanced ratio between correct and incorrect language use within the solution of the student.

Provided a correct and covering model can be supplied for a limited domain, diagnosis is guaranteed to be precise and robust enough and error anticipation eventually may be renounced completely. In order to yield an efficient implementation of the idea into a practical solution a preponderantly data driven procedure instead of a strictly hypotheses driven one seems to be desirable. A procedure of this kind has been successfully pursued in an earlier paper on the diagnosis of agreement errors in fixed syntactic environments (Menzel 1988). Quite naturally this success raises the question on how much of the experience gathered can be transferred to other types of grammatical regularities as linear ordering principles or dominance regularities, for instance.

Up to now the only notable exception to the one-sided orientation on error anticipation has been a fail-soft technique implemented in the error sensitive parsing system Linger (Barchan et al. 1986), an approach which later has been named "word soup heuristics" by their authors: Whenever the normal parsing process based on a principally anticipation-oriented context-free grammar fails, the system attempts to achieve a successful parse by trying single word form substitutions, insertions, deletions or displacements. Although often being very useful in detecting simple flaws of the student, this heuristics not so infrequently produces rather surprising and sometimes even funny interpretations of the input data. Its main drawback is the basic limitation to only single word form errors. Any extension to the handling of complete constituents, desirable as it may be, seems to be condemned to failure because of efficiency reasons: the whole approach is basically expectation driven and it opens up too vast a search space of possible error hypotheses, where the verification of only a single one is not just a trivial task.

I. MODEL-BASED DIAGNOSIS

The intrinsic problem with the diagnosis of structural errors is its not fitting easily into the standard paradigm of model-based reasoning which essentially relies on two basic assumptions (Reiter 1987):

(1) A model always has an a priori given number of elementary model components.
(2) The intercomponent connections of the model are invariant and, likewise, given a priori.

Accordingly, model-based diagnosis primarily should be applicable to domains with a fixed and known structure, which is typical for e.g. electronic troubleshooting, the origin of the approach. Provision can be taken for these premises to be fulfilled in artificially limited domains of natural language, e.g. for agreement errors (c.f. Menzel 1988). In more natural environments of language production, however, they do not hold. Parsing a natural language sentence first of all is solving the task of structural identification. Therefore, diagnosing arbitrary syntactic errors in arbitrary utterances, from this point of view, may perhaps be compared with electronic troubleshooting in a circuitry of obscure function with at least partially unknown components and partially invisible wiring under the additional assumption that there is no possibility to vary the conditions of measurement! It should go without saying that such a task can only be solved in very limited domains relying on an as strong as possible (semantic) support from the situational and sentential context. The final goal, of course, should be an as far-reaching as possible integration of structural identification and diagnosis.

Model-based diagnosis, especially for teaching purposes where comprehensible error explanations are desired, poses two additional constraints on the kind of model information to be used. Both conditions, if compared against usual parsing grammars, certainly are not a matter of course:

(3) The model has to provide an extremely reliable correct/incorrect distinction, whereas traditional grammars, in the hope that ungrammatical sentences will not appear as input, massively rely on overgeneration.
(4) An explicit representation based on comprehensible generalizations has to be

attempted for a maximum of regularities in the domain, in order to allow this information to be used immediately for explanation purposes.

The latter condition in most cases definitely rules out simple lists of alternative solutions as a proper means of representing model information. To code, for instance, word order regularities as a list of possible permutations gives no sensible error explanation besides, say, "Your constituent order is not contained in the list of admissible constituent sequences". What is desired instead of this would be an explanation, based on explicit generalizations as in "The verbal group of German subordinate clauses has always to be placed in final position".

A first attempt to make word order regularities explicit has been made by using ID/LP format for GPSG (Gazdar et al. 1985). For diagnostic purposes such explicitness is not only necessary with respect to linear order principles but with respect to omissibility and combinability of categories as well. It should result in a clear distinction between a rather simple notion of dominance rules and a comparatively rich set of various constraints over dominance structures.

II. DOMINANCE STRUCTURES

In its most simple case a dominance rule

$A \longrightarrow B_1, B_2, \ldots, B_n$

states the ability of category A to dominate all sequences of categories which are arbitrary permutations of list $L=(B_1, B_2, \ldots, B_n)$ or of any not empty sublist of L. According to this definition, a dominance rule can easily be interpreted as a disjunction of elementary and independent (local) dominance relations dom[X, Y] and each category in list L represents an optional constituent:

or[dom[A, B_1], dom[A, B_2], ..., dom[A, B_n]],

or in a shorthand notation:

or[B_1, B_2, ..., B_n].

If a highly precise and explicit representation of dominance regularities (according to condition (3) and (4)) is aimed at, this simple rule format is obviously not sufficient. It does not even allow the usual distinction between optional and obligatory elements in the list of dominated nodes, and especially for the purpose of model-based diagnosis a further refinement is inevitable.

Obviously, a minimal formal base should contain at least an explicit description of the sometimes rather intricate compatibility conditions between elementary dominance relations, e.g. by means of propositional expressions. In order to yield a simple diagnosis procedure, the complexity of admissible expressions has to be carefully restricted. For a good number of cases a conjunctive combination of elementary (usually binary) expressions is already sufficient. Such elementary expressions then can be interpreted as additional constraints for the simultaneous appearence of categories within a constituent in a very similar way as agreement and word order constraints restrict the compatibility of inflected forms or the sequencing of categories.

Most often needed are compatibility constraints to describe an alternative (exor), an implication (if) or an equivalence (iff) of dominance relations. Additionally, a dominance relation can be made obligatory, if it is simply specified as a single element in the conjunction of constraints. Hence, by choosing a sensible specification of constraints optionality or obligatoriness can easily be expressed as special cases. In the simple noun phrase

NP ⟶ Det, Adj, Noun

the determiner and the noun can be indicated as obligatory by adding the constraints:

and[Det, Noun]

whereby the adjective remains optional.

A more ambitious example could be the a German local prepositional phrase PP

PP ⟶ Prep-3, Prep-3-Det, Det, Adv, Adj, Noun

which allows in addition to the usual dative prepositions (Prep-3) the fusion of preposition and determiner (Prep-3-Det) which is very common not only in spoken German ("an" + "dem" = "am", "in" + "das" = "ins", etc.). The additional constraints

and[exor[Prep-3, Prep-3-Det], or[Adj, Noun],
iff[Prep-3, Det], if[Adv, Adj]]

provide for a prepositional phrase to contain one and only one preposition and exactly one determiner, independently of being fused or not. Adjective and noun both are optional (but not simultaneously) and the admissibility of an attributive adverb depends on the existence of the modified adjective.

To describe the omissibility of dominated nodes (e.g. for the determiner) arbitrary elementary logical conditions (e.g. for the presence or absence of certain semantic features) may be included into the set of constraints.

This simple formal framework certainly is not sufficient to write complex grammars. Nevertheless, it can serve to build tiny (but non-trivial) specialized grammars covering e.g. simple types of main or subordinate clauses, extended noun phrases with left and/or right attributes etc. which then meet the rather strong preconditions for an application of model-based diagnosis techniques.

III. DIAGNOSIS

According to the strong bias within the descriptive framework towards consistency constraints, the bulk of student errors will have to be diagnosed as consistency violations. For that purpose a constraint propagation procedure based on constraint retraction or, logically stronger, constraint negation can be used. It is this kind of procedure by which agreement errors earlier have been tackled successfully. Now it turns out that linear ordering principles can be handled in a quite similar way. The only serious distinction is the origin of factual information: Whereas for agreement it is taken from the dictionary (morpho-syntactic features), for word order it is given as a position number in the input sequence chosen by the student.

Mutual constituent incompatibility and the omission of obligatory constituents are diagnosed as violations of the above mentioned constraints. Constraint negation as basic diagnostic technique is a comparatively simple procedure in the case of e.g. disjunction and implication. In both cases the reason for the constraint violation is unique. More attention requires e.g. the violation of an alternative where two cases (which result in two different explanation variants) have to be properly distinguished: None of the required categories has been detected vs. both categories appear simultaneously.

Combinatorial problems arise out of the transitivity property of some constraints. This is typical for agreement constraints and in most cases it rules out a local decision upon a particular error hypothesis. Additional difficulties arise out of (legal or illegal) constituent omissions, where constraint propagation has to take into consideration a (locally limited) transitive closure of constraint relations.

IV. PARSING AS CONSTRAINT SATISFACTION?

Shifting information from traditional syntactic rules into additional constraints makes parsing an increasingly difficult enterprise. Valuable information usually used to reduce the search space has been lost.

With the basic definition of dominance to be a disjunctive combination of potential dominance relations a grammar can be interpreted as an OR-tree and structural identification (parsing) becomes a procedure of attaching all the categories occurring in the input sequence to corresponding leaves of the grammar tree. This, of course, restricts practical solutions to finite trees, i.e. nonrecursive dominance relations. For nonrecursive relations the search space becomes finite but remains nevertheless extremely large. Even for very small grammars many combinations of category attachments exist, each of which stands for a separate constraint satisfaction problem, which to solve, under normal circumstances, again requires a combinatorial procedure.

Hence, a further drastic reduction of the search space has to be achieved by means of different heuristics:

1. Certain compatibility constraints which are unlikely to be violated by the student (e.g. two prepositions in a single PP) can be made implicit. In that case, they cannot be violated and consequently not be used for explanatory purposes. They can, however, be well used to exclude senseless category attachments.

   Since this heuristics often applies for clearly alternative dominance relations, the tree of dominance possibilities gets an implicit or/exor-structure. A very similar technique can be applied for optional subordinated constituents, where the subordinated constituent should be accepted only, if at least something of of the dominating constituent has been identified.
2. Exclusion of useless permutations during category attachment to the grammar tree by maximising a locality measure. This heuristics fails systematically in case of certain embedded constituents, where it prefers to attach e.g. a determiner to the embedded noun phrase instead of assigning it to the more distant noun.
3. Reuse of partial results guided by a clustering of functionally different constituents according to their structural equivalence, e.g. NPs, PPs etc.
4. Data driven best first analysis.

For simple grammars (of the above mentioned complexity) these heuristics usually reduce the ambiguity of attachment to only a few readings which remain to be passed to the consistency check. Since even very simple grammars often are quite sufficient for language learning purposes, indeed a kind of restricted, but nevertheless useful parsing system mainly based on constraint satisfaction techniques can be devised. This at least allows to considerably soften the strong limitations of the approach in Menzel (1988), which are imposed by the restriction to fixed syntactic environments. Considering however the rapidly growing search space required for more complex models, an extension of the approach to the level of a universal grammar obviously is not feasible.

As a result of the limits of a simple model a few types of errors cannot be diagnosed as constraint violations but have to be detected already during the procedure of category attachment. In particular this concerns the detection of superfluous forms e.g. the use of two finite verbs within a single sentence. But generally the preference of an insertion as error hypothesis is rather low. The suspicion of misinterpreting the student's intentions should be much more justified. Such substitutions of categories result in a combination of a category insertion and omission at the same place in the utterance. If a single word form is concerned this constellation sometimes indicates a wrong application of word formation rules or inflectional schemes which result in an unintended category. E.g. "kochen" (to cook) is a finite verb but in the sentence "*Die Kochen gehen nach Hause" (approximately: *The cooking are going home) it has to be interpreted as a mistaken plural of "Koch" (the cook). Diagnosis infers this from the missing noun of the subject and the superfluous verb, supported by the unusual capitalization of the supposed verb.

Barchan, J; Woodmansee, B. and Yazdani, M. (1986) A PROLOG-based Tool for French Grammar Analysis. in: Instructional Science, vol. 14, p. 21-48.

Catt, M. E. (1988) Intelligent Diagnosis of Ungrammaticality in Computer-Assisted Language Instruction, Technical Report CSRI-218, Computer Systems Research Institute, University of Toronto.

Gazdar, G., Klein, E., Pullum, G. K., Sag, I. A. (1985) Generalized Phrase Structure Grammar, Oxford.

Menzel, W. (1988) Error Diagnosing and Selection in a Training System for Second Language Learning, Proc. 12th Coling 88, Budapest: 414-419.

Reiter, R. (1987) A Theory of Diagnosis from First Principles, Artificial Intelligence, vol. 32, no. 1: 57-95.

Schwind, C. B. (1988) Sensitive Parsing: Error Analysis and Explanation in an Intelligent Language Tutoring System, Proc. Coling 88, Budapest: 608-613.

Yazdani, M. (1988) Language Tutoring with PROLOG. Papers of the International Workshop on Intelligent Tutoring Systems for Second Language Learning, Trieste: 150-155.