Parsers Optimization for Wide-coverage Unification-Based Grammars using the Restriction Technique

Nora La Serna Arantxa Díaz
Department of Computer Languages and Systems, University of the Basque Country
p.k. 649, 20080 Donostia, Spain. (jiblapan@si.ehu.es)

Horacio Rodríguez
Department of Computer Languages and Systems, Universitat Politécnica de Catalunya
Pau Gargallo 5, 08028 Barcelona, Spain. (horacio@lsi.upc.es)

Abstract

This article describes the methodology we have followed in order to improve the efficiency of a parsing algorithm for wide-coverage unification-based grammars. The technique used is the restriction technique (Shieber 85), which has been recognized as an important operation to obtain efficient parsers for unification-based grammars. The main objective of the research is how to choose appropriate restrictors for using the restriction technique. We have developed a statistical model for selecting restrictors. Several experiments have been done in order to characterise those restrictors.

1. Background and Tools used

The use of linguistic material associated to the features of unification-based grammars, for guiding the parsing process may give a problem, as Shieber pointed out in (Shieber 85). This is, certain unification grammars present an infinite nonterminal domain, leading to inefficiency or even non-termination of the algorithms if standard methods of parsing are used. At the same time, Shieber proposed a solution to this problem with the *restriction* technique. However, there is an open question using restriction. It is not clear how to choose an appropriate *restrictor* (subset of the feature structures owned by the complex categories of the grammar), which assures to obtain the greatest efficiency. An inadequate choice of restrictors affects the efficiency of parsing algorithm. The research presented in this article deals with this problem i.e, what we present here is a methodology for choosing adequate restrictors with wide-coverage unification-based grammars.

In our study, a Patr-II grammar has been generated from the object grammar of the system ANLT (Alvey Natural Language Tools; Grover et al. 93 & Carroll 93). The Alvey grammar defines a wide-coverage of syntactic grammatical constructions of English; only 350 rules and 5008 entries of the grammar and lexicon respectively have been converted. We have also used the UNICORN parser developed by Gerdemann and Hinrichs (Gerdemann 91). It is an Earley-style chart-parser for unification grammars that incorporates the restriction technique.

Restriction is defined as follows: Let R be a *restrictor*, and D a dag of a unification grammar. RD is the *restricted dag* for D relative to R if RD subsumes D, and for every path p in RD, there exists a path q in R such that p is a prefix of q.

2. Methodology and Experiments Developed for Selecting Restrictors

The method we have used for selecting adequate restrictors with the wide-coverage grammar is based on the criterion of instantiation of the features. So, a statistical model for estimating the probabilities of being instantiated of the different features of the grammar along a parsing process was defined. Two important reasons justify the model: 1) each rule of the grammar has different application probability; 2) instantiation can be achieved by appearing explicitly

instantiated in the rule or by percolating through mother or sister categories. Formally, we present the model as follows:

$$Pinst\ (C_i, R_j) = \sum_k \Big(P(r_k) * (instantiated\ (C_i, R_j, r_k) + \Big(P(Cm_k, R_j) + \sum_p P(Ch_k, R_j) \Big) * un\_instantiated\ (C_i, R_j, r_k) \Big)$$

In the formula, Pinst (C_i, R_j) is the probability that the feature j, of the category i, appears instantiated; r_k is the kth rule; Cm is the mother category; and Ch is the brother category. For each rule of the grammar where the feature j, of the category i is defined, the following can be done: 1) If the feature mentioned *is instantiated*, the matrix *instantiated* (C_i, R_j, r_k) will contain the value one, and zero otherwise; 2) If the feature mentioned *is uninstantiated*, the matrix *no_instantiated* (C_i, R_j, r_k) will contain the value one, and zero otherwise. In this step, the value of the feature can be inherited from the mother or brother categories. 3) If the feature analyzed is uninstantiated, and the value of the feature can not be inherited, then the Pinst (C_i, R_j) is zero in the rule. Finally, the model generates a regular linear equations system, where the variables are the probabilities of being instantiated.

In order to prove our criterion for choosing appropriate restrictors we have performed several types of experiments (La Serna 96). The aim of them was to observe which of the proposed restrictors are adequate. The experiments consist of the analysis of a set of 30 phrases (selected randomly from the corpus) with different restrictors, basically change in length and grades of instantiation (high, intermediate, and uninstantiated). The selected measure of evaluation is the number of predictive states, so the *apropriate restrictors* are those which have the lowest number of states.

The experiments have been planned with two classes of restrictors: *static* and *dynamic*. In the first class, the restrictor is the same for all the categories of the grammar, as established in the original definition of restriction. In the second class, we have proposed that the restrictor can be different for each category of the grammar, because certain features make up adequate restrictors in some categories, but are not good candidates for others.

From the analysis of the results of all the performed experiments, we can point out the following: 1) Features that have established *appropriate restrictors*, which are instantiated in all the grammar rules. For restrictors with uninstantiated features in at least some rule, the results have not been better, in general. 2) Any combination of features of appropriate restrictors form again appropriate restrictors; however, any combination of non-appropiate restrictors makes non-efficiency process parsing. 3) In the experiments with static restrictors, the features established as appropriate restrictors have been obtained from the intersection of the best features of each category in the dynamic restrictors. 4) Finally, we observed that with the dynamic restrictors, there are more possibilities for choosing appropriate restrictors.

References

(Carroll 93) Carroll J. (1993). *Practical Unification-Based Parsing of Natural Language*. PhD thesis, Cambridge University,UK.

(Gerdemann 91) Gerdemann D. (1991). *Parsing and Generation of Unification Grammars*. PhD Thesis, University of Illinois, Technical Report CS-91-06 of The Beckman Institute.

(Grover et al. 93) Grover C., Carroll J., Briscoe T. (1993). *The Alvey Natural Language Tools Grammars(4th release)*. Technical Report N° 284, Computer Laboratory, Cambridge University,UK.

(La Serna 96) La Serna N. (1996). *Selecting Apropriate Restrictors with Wide-Coverage Unification-Based Grammars*. Research report UPV/EHU/LSI/TR15-96, University of the Basque Country, Spain.

(Shieber 85) Shieber S. (1985). *Using Restriccion to Extend Parsing Algorithms for Complex Feature Based Formalisms*. In ACL Proceedings, 23rd Annual Meeting, University of Chicago. Chicago, ILL.