Computational Aspects of Constraint-based Linguistic Description II

Jochen Dörre
(editor)

DYANA-2

Dynamic Interpretation of Natural Language
ESPRIT Basic Research Project 6852

Deliverable R1.2.B
September 1994

Computational Aspects of Constraint-based Linguistic Description II

Jochen Dörre
Universität Stuttgart
(editor)

authors:
Andreas Eisele
Chris Brew
Steve Hegner
Chris Mellish
Jochen Dörre, Dov Gabbay and Esther König
Hans Leiß

DYANA-2

Dynamic Interpretation of Natural Language
ESPRIT Basic Research Project 6852
Deliverable R1.2.B
September 1994

ESPRIT Basic Research Project 6852 DYANA-2

Dynamic Interpretation of Natural Language

Universiteit van Amsterdam
Institute for Logic, Language and Computation (ILLC)

University of Edinburgh
Centre for Cognitive Science (ECCS)

Universität München
Centrum für Informations- und Sprachforschung (CIS)

Universitetet i Oslo
Department of Linguistics and Philosophy (ILF)

Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung (IMS)

Universität Tübingen
Seminar für Sprachwissenschaft (SfS)

Universiteit Utrecht
Research Institute for Language and Speech (OTS)

For copies of reports, updates on project activities and other DYANA-related information, contact:

The DYANA-2 Project Administrator
ILLC/Department of Philosophy
University of Amsterdam
Nieuwe Doelenstraat 15
NL-1012 CP Amsterdam

Contents

Introduction

This collection consists of three papers on issues in constraint-based grammar formalisms, as well as a commentary on each of these. All three report on research which is strongly relevant to the development of CUF, DYANA's grammar specification formalism, however no preknowledge of CUF is actually required.

The first paper (Eisele: "Towards Probabilistic Extensions of Constraint-Based Grammars") presents an approach which attempts to tame the power of (grammar) specification formalisms like CUF by adding probabilistic control information, which can be used to guide the search. The second paper (Hegner: "Distributivity in Incompletely Specified Type Hierarchies") focuses on the new type specification language of CUF and presents a thorough algebraic treatment of extension problems that come with such specifications facilitating the comparison to other such languages, for instance, to the type language of ALE. Finally, the contribution by Dörre, Gabbay and König on "Fibred Semantics for Feature-based Grammar Logic" uses the new fibred semantics paradigm to build a formal semantics for a class of categorial-based unification grammars. Implicit to any of these contributions is the theme of taming or restricting the power of a general specification language, such as CUF. Let us present the papers now in more detail.

Probabilistic CUF Andreas Eisele draws the detailed layout of a probabilistic extension of CUF. This is highly innovative work — although maybe not unexpected when considering the current massive trend towards statistical methods — combining a stochastic approach with a symbolic approach with the latter being much more expressive than in other such combinations which can be found in the literature.

Eisele chooses to assign probabilities to the clauses of a CUF program thereby turning the nondeterministic search into a stochastic procedure which, seen abstractly, 'outputs' proofs with certain probabilities. Actually, this move may also be viewed as taking the notion of control information to the extreme where the chance of arriving at a global solution when choosing a local option, so to speak the quality of the option, becomes quantifiable and search may be completely guided by a regime that always selects the path with the best estimate for a global solution. One variant of such a best-first search which is based on generalized Earley deduction is sketched. This strategy enables us to find optimal solutions in polynomial time even in certain cases where simpler strategies have to investigate a search space of exponential size.

The approach taken allows to induce maximum likelihood estimates for the probabilistic parameters in a given description from *training examples*. A method for achieving this is described, which is a variant of the EM-algorithm, the method of choice for the training of Hidden Markov Models and stochastic context-free grammars. However, the interaction between constraints on features and probabilities attached to rules raise some questions which could not yet been answered conclusively, and which require further theoretical and practical investigations.

The contribution also discusses some smoothing methods that address the difficulty of getting accurate estimates from limited training data and shows how a smoothed, class-based model can be expressed in probabilistic CUF.

Note that this approach is quite different from the proposal made by Chris Brew in last year's deliverable, where probabilities are associated with types using features and special numerical built-in predicates are required to explicitly program probabilistic procedures. In a comment to Eisele's paper, Brew advocates again this simpler, yet more limited approach and points to possible disadvantages that some of Eisele's decisions may have in practice.

Type specifications & Algebra The motivation in Hegner's contribution is the wish to make algebraic sense of CUF's powerful and hence computationally problematic type specification language and thus to facilitate the comparison to other such languages with algebraic semantics, especially the one of the ALE formalism. Stated more generally, he investigates the problems involved when having to extend a partially specified type hierarchy (given as some poset), for which also some greatest common subtypes and least common supertypes (as equations $t = t_1 \wedge \ldots \wedge t_n$, resp. $t = t_1 \vee \ldots \vee t_n$) are known, to a distributive lattice, i.e., one in which indeed the intuitive interpretation of subtyping as set inclusion, $\wedge$ as intersection and $\vee$ as union can be assumed. In order to systematically study different extension problems, Hegner sets up a rigorous mathematical framework of the theory of bounded posets with partial operations (BPPOs) and their extension morphisms in the categories of three different kinds of distributive lattices: the bounded distributive lattices, the bounded Boolean lattices, and the so-called perfect Boolean lattices. This framework provides us with a tool for the precise formulation of the respective variations of extension problems as well as for studying their systematic relation.

In a second part, Hegner studies the computational complexity of these problems with special attention to their dependence on the parameters *height* (the length of the longest path in the specification), *join fanout* and *meet fanout* (the number of elements combined with a join, resp. meet, operation) of the input BPPO prespecification. In the reductions the notion of separability of distinct types, i.e., the question whether there are models distinguishing them, plays a crucial role. Here is a short summary of the main results.

- Even under very tight constraints concerning the input parameters (one of the three parameters equal to 3 and the other two equal to 2) all of the extension decision problems are $\mathcal{NP}$-complete.
- If all parameters are equal to 2, i.e., only for relatively trivial prespecifications, the problems can be solved in deterministic polynomial time.
- The extension problem also allows for a deterministic polynomial time algorithm, if we admit only join or only meet restrictions. The latter, for instance, is the case for ALE's rather weak type specification language.
- When we also take into account the declaration of atoms (like constants in CUF) the property that a BPPO always has a free extension, i.e. a natural one which imposes the minimum set of constraints possible, gets lost. These declarations appear to be responsible for an additional source

of nondeterminism.

Fibred Semantics The last contribution, by König, Gabbay and myself, is concerned with the formal semantics of grammar formalisms based on the combination of categorial-style grammars, the Lambek calculus being the paradigm example, and feature logic. For this combination Gabbay's method of fibred semantics is used which amounts to syntactically allowing formulas of one logic in the places of atoms in formulas of the other and semantically associating models of one logic with elements in a model of the other via so-called fibring functions. Two variants of the combination are considered.

In the first basic categorial types b are replaced by formulas $b(T)$ where T is a feature term. However, the logic that one gets when employing a fibred semantics straightforwardly does not seem to be the underlying logic of a proof system which is like Lambek's, but uses unification (of feature graphs) for the matching of categories, like in the formalism of categorial unification grammar which the paper intends to cover.

Therefore a second mode of combination is considered in which only variables are attached to the basic categories and feature descriptions are employed as *global constraints* on these variables. For this logic the anticipated proof theory (Lambek + unification) turns out to be appropriate. However, since in the intended use of the grammar formalisms in question the parsing problem is not mapped to a mere problem of derivability, but requires the collection of consistent constraints that license a derivable sequence, we are consequently not just interested in the mere problem of validity or consequence, but rather in the set of solution constraints of a goal. The problem of computing solution constraints actually contains a mixture of a satisfiability problem (solution constraints need to be satisfiable) and a validity (or consequence) problem (in the class of models of the solution constraint the goal must be valid). The inherent modularity of the fibred semantics approach facilitates the definition of this notion of solution.

The main result of the paper is a proof theory which is sound and complete w.r.t. *generating answers* for the second kind of combined logic and which represents a more abstract characterization of the intended proof theory, Lambek calculus augmented with unification. An examination of the computational complexity reveals that the problem of determining whether there exist solutions to a given goal is $\mathcal{NP}$-complete even in restricted contexts, where neither the Lambek part nor the feature constraint part are responsible for $\mathcal{NP}$-completeness on their own.

An aspect of central importance is that in the favored second combination scheme the proof systems of both component logics carry over essentially unchanged. Moreover, the approach is actually much more general than is suggested in the previous text, since this combination scheme does not rely on any specific property of the grammar logic besides that there must be a notion of basic category to which we can attach the variables.

This allows us to combine virtually any (propositional) categorial logic in a systematic way with a constraint logic. Note that such an integration of the categorial into a sign-based approach is also desirable from a computational point of view. The categorial part takes care of functor–argument structure ('subcategorization' in HPSG terms) and of issues of unbounded dependencies in a coarse manner, whereas constraints provide a particular simple way to define refinements of these dependencies. Hence, the most complex parts of HPSG signs are pushed into a propositional framework, for which parsing can be done much more efficiently.

The collection finally includes a comment of the last paper by Hans Leiss pointing out some deficiencies in the presentation and some (maybe) unnecessary complications in the formal parts. Unfortunately time constraints did not permit to take up all of his suggestions for improvement. As one of the authors I am grateful to him for the detailed and constructive criticism, although I would like to add that some of his questions seem to arise from a misunderstanding of our primary goal, namely to define a formal semantics of an implemented grammar formalism.

Stuttgart, September 1994
Jochen Dörre

Task 1.2, subtask 1

Computational Aspects of Constraint-based Linguistic Description

Towards Probabilistic Extensions of Constraint-Based Grammars

Andreas Eisele
(Universität Stuttgart)

Types and Clauses: Two styles of probabilistic processing in CUF

Chris Brew
(University of Edinburgh)

Additional Contribution

Distributivity in Incompletely Specified Type Hierarchies: Theory and Computational Complexity

Steve Hegner
(Universitetet i Oslo)

Comments on "Distributivity in Incompletely Specified Type Hierarchies"

Chris Mellish
(University of Edinburgh)

Additional Contribution

Fibred Semantics for Feature-Based Grammar Logic

Jochen Dörre, Dov Gabbay and Esther König

(Universität Stuttgart, Imperial College (London), Universität Stuttgart)

Comments on "Fibred Semantics for Feature-Based Grammar Logic"

Hans Leiß
(Universität München)