

Automating Reuse of Broad-Coverage Typed Feature Structure Grammars

1. Introduction

This paper presents research ideas of inducing typed feature structure grammars (TFSG) by reusing broad-coverage TFSGs of related languages. The devised methodology combines symbolic, probabilistic, and empirical methods in order to achieve its goal.

2. Background

Since the beginning of the Chomskyan tradition, grammar writers and implementors have taken up the challenge to produce as accurate and covering monolingual grammars as possible. Often, this has taken place within certain theoretical and practical frameworks, more or less relating to existing linguistic theories, sometimes even contributing to language theory.

For computational applications, the obvious and natural strife to achieve broad coverage and use representations which facilitate efficient implementations has to some extent overshadowed other important aspects of monolingual grammars.

The question of how much of a computational grammar is parochial, and how much belongs to a universal “core” is often overlooked.

Although our scientific communities may agree on a certain collection of formal elements which constitute typed feature structure grammars (TFSG) (subsumption relations of type hierarchies, appropriateness conditions), and this collection of formal elements would prove adequate to describe relevant properties of every single language in the world, it is not given that TFSG constitutes an adequate form of representation in order to distinguish a common core from the rest of the formal elements. On the one hand, TFSG may not be fine-grained enough. On the other hand, the core grammar might be expressed as the combination of a huge type hierarchy allowing every possible kind of multiple inheritance, and a large disjunction of language-specific constraints, leaving it as a mathematical task to discover cross-linguistic generalizations by factoring out repeating terms.

3. Broad-coverage grammars of related languages

Development of formal and implemented broad-coverage grammars in TFSG frameworks (and other frameworks indeed!) has proven a difficult and time-consuming task for even trained linguists. Although it is of great interest of the scientific community to compare the degree of diversity or uniformity of results of “competing” efforts on related languages or related phenomena within languages differing in other respects, there are other considerations to make. For practical purposes, it is tempting to use a TFSG covering one language as the starting point for the development of a TFSG for a related language.¹

4. Automatically reusing TFSGs of related languages

The author has committed himself to shed light onto the following question:

Can we, given a TFSG-account of one language(’s morphology and syntax) (L1), automatically induce a probabilistic grammar of another language (L2)? This involves the treatment of two subquestions:

1. How much do the constraints of L1 have to be relaxed to yield an over-generating binary TFSG grammar which allows a lot of language structures - including L2 (and also L1)?

¹In practice, this is what happens when a group of researchers discovers a well-known phenomenon in a new language. It is of course utterly bad practice to reinvent the wheel. You apply your inventory of standard analyses as far as possible, only renouncing when a convincing generalization, simplification, or external explanatory device applies.

2. Can this grammar (in a very special sense - one candidate core grammar for a language pair) be used as a basis of learning/inducing corresponding probabilistic grammar versions respectively preferring sentences of L2 when trained on a L2 corpus — and L1 sentences when trained on a L1 corpus?

The author envisages the following methodology:

1. Apply as resource the following:
An existing TFSG for L1 (e.g.; English: Lingo; Norwegian: NorSource; German: Babel);
Corpora for L1 and L2; Bilingual wordlists for L1 and L2 (translation lexicon);
2. For a chosen set of constraints C of $G(L1)$, establish a generalized probabilistic constraint C_g (or a set of such). This is made parsable in L2 by some bootstrapping penalty - if possible.
3. Use existing learning methods (Riezler, Johnson) combined with bilingual wordlists in order to optimize the penalties of C_g .
4. Iterate through the powerset of C , thus ensuring that the L2 grammar is not more general than necessary.

To avoid the trap of solving many unrelated issues at once, the initial objects of study (constructions) should be phenomena which are well-described in both languages, in order to judge the outcome of the experiments.

5. Conclusions, prospects, and status

Positive results of this undertaking might drastically reduce the amount of resources needed to develop broad-coverage grammars for related languages.

It is at present unknown what status in the inventory of theoretical linguistic concepts a common binary (“traditional”) TFSG of two related languages has. Such may be postulated for whatever reason. The present project may shed light on this question by giving a measure of distance between pairs of language fragments as a metaphor of “formal probabilistic learnability” from a common source.

The adaptation of learning algorithms as well as conducting pilot experiments are ongoing.

6. *

References

- [1] Mark Johnson. Learning and parsing stochastic unification-based grammars. In *COLT*, pages 671–683, 2003.
- [2] Carl J. Pollard and Ivan A. Sag. *Head-Driven Phrase Structure Grammar*. University of Chicago Press, Chicago, 1994.
- [3] Stefan Riezler. Learning log-linear models on constraint-based grammars for disambiguation. In James Cussens and Saso Dzeroski, editors, *Learning Language in Logic*, volume 1925 of *LNCS*, pages 199–217. Springer Verlag, June 2000.
- [4] Peter Rossen Skadhauge. Implementation of a typed feature structure grammar processing system with a probabilistic processing component. *Acta Linguistica Hafniensia*, 36, 2003.