Improving Treebank-Based Automatic LFG Induction for Spanish

Grzegorz Chrupała and Josef van Genabith
NCLT, School of Computing, Dublin City University
grzegorz.chrupala@computing.dcu.ie
josef@computing.dcu.ie

Introduction

In the present paper we report on ongoing work on automatic induction of wide-coverage Lexical-Functional Grammar resources from the Spanish part of the 3LB treebank (Civit and Martí, 2004). A methodology for automatically obtaining LFG f-structures from trees output by probabilistic parsers trained on the Penn II treebank has been described by (Cahill et al., 2004). It has also been shown that the same methods can be ported to other languages and treebanks (Burke et al., 2004; Cahill et al., 2003), including Cast3LB (O’Donovan et al., 2005).

Here we describe improvements made to the methods used by (O’Donovan et al., 2005) and report the results of evaluation of constituent parsing, Cast3LB functional label assignment and f-structure construction.

The enhancements described in the current paper fall into two categories. Firstly, we use machine-learning methods to assign functional labels to nodes in parser output to help in obtaining higher-quality f-structures. Secondly, we expand our coverage of Spanish construction types and linguistic phenomena.

Grammatical function assignment

Some properties of Spanish and the encoding of syntactic information in Cast3LB treebank make it non-trivial to apply the method of automatically mapping c-structures to f-structures used by (Cahill et al., 2004), which assigns grammatical functions to tree nodes based on their phrasal category, the category of the mother node and their position relative to the local head.

The order of sentence constituents is flexible and their position relative to head is an imperfect predictor of grammatical function. Also, much of the information that the Penn-II Treebank encodes as tree configurations is encoded in Cast3LB in the form of function tags.

Thus, assigning correct LFG functional annotations to nodes in Cast3LB-style trees is difficult without use of function tags, and those tags are typically absent in probabilistic parsers’ output.

In order to leverage functional tag information, (O’Donovan et al., 2005) train the parser to output complex category-function labels. This, however, inflates the number of unique labels and can introduce sparse data problems.

The approach we adopt instead is to add Cast3LB functional tags to constituent tree parser output as a postprocessing step, following (Blaheta and Charniak, 2000), prior to f-structure annotation. We experimented with three machine learning algorithms: Memory-Based Learning, Maximum Entropy and Support Vector Machines. For all the algorithms we extract the same set of features encoding configurational, morphological and lexical information for the target node and neighboring context nodes. For each of the training examples we also extract the Cast3LB function tag, or assign the default null tag if no tag is present.

In order to evaluate the performance of the classifiers trained using our approach, we measured the f-score achieved on sets of tuples of the form \((GF, i, j)\), where \(GF\) is the Cast3LB grammatical function tag and \(i – j\) is the range of tokens spanned by the target node. With the best-performing classifier, SVM, our f-score improved by 6.74% in comparison to the baseline parser-based method used by (O’Donovan et al., 2005).

We also evaluated the quality of the f-structures obtained by annotating the trees with Cast3LB function tags assigned by the SVM classifier. We convert the f-structures to dependency triples and compare against a gold standard of 338 manually corrected f-structures. We improve the f-score by 3.4% over the baseline.

Grammar coverage

There are two very common features of Spanish whose account in LFG involves the use of optional functional annotations. In order to construct correct f-structures for sentences where those features occur we implement optionality in the constraint solver used to evaluate functional equations.

Spanish is a null subject language – explicit subjects are often absent. Subject features such as person and number are encoded in agreement morphology on the verb instead. Following standard
LFG practice we treat this phenomenon by including an optional subject specification in lexical rules for verbs: \( ((f \ subj pred) = 'pro') \).

Following the proposal by (Andrews, 1990), optional descriptions are also used in treating another common feature of Spanish, i.e. clitic doubling. Spanish verbal clitics expressing direct object and indirect object can co-occur with NPs expressing the same grammatical function (this is especially common for indirect objects). To account for this we add appropriate equations to lexical entries for pronominal clitics: \( ((f pred) = 'pro') \).

We also enhance our annotation algorithm in order to build correct f-structures for certain types of long-distance dependencies, namely relative clauses and direct and embedded questions. Our treatment in not yet fully general: we do not currently deal with dependencies of unbounded length due to limitations in the 3LB annotation scheme and the small number of sentences with such dependencies. For the most frequent monoclusal case, however, our implementation correctly constructs f-structures with the grammaticized discourse functions \( \text{topic} \) and \( \text{focus} \).

A difficult issue in LFG treatments of Romance languages is the verb (or periphrastic) constructions, where two or more verbs share a set of grammatical functions and form a complex predicate (Andrews and Manning, 1999). There is no agreement on what an LFG account of complex predicates should look like – most proposals (Alsina, 1997; Butt, 1997; Andrews and Manning, 1999) involve extensions to standard LFG.

In our computational implementation we have decided to avoid non-standard extensions and conservatively deal with Spanish periphrastic constructions in terms of nested XCOMPS. We added functional uncertainty and off-path constraint support to our constraint solver in order to achieve this. In the general case, those devices are necessary to account for clitic climbing, where clitics expressing an argument of a lower verb attach to a higher verb. Additionally, in the case of Cast3LB they are also necessary for ordinary NP arguments due to the idiosyncratic encoding of periphrastic constructions in the treebank.

Conclusions

We have described adaptations to the automatic f-structure annotation made in order to accommodate features of the Spanish language and of the representation of syntactic structure in the Cast3LB treebank, which let us obtain high-quality f-structures. We have also reported on the expansion of coverage of linguistic phenomena found in Spanish. The solutions we used should also prove useful in porting our data-driven LFG induction approach to other Romance language treebanks, such as the Catalan version of 3LB.

References


