Procesamiento del Lenguaje Natural, num. 30 (2003), pp.123-124 recibido 1-3-2003; aceptado 20-5-2003

An Island-Driven Parsing System*

Alicia Ageno Pulido
Director: Horacio Rodríguez Hontoria
Universitat Politècnica de Catalunya (Dept. LSI)
Jordi Girona, 1-3. 08034 Barcelona
ageno@lsi.upc.es

Resumen: El núcleo de esta tesis presenta y evalúa dos métodos para modelizar probabilísticamente la bidireccionalidad en el análisis sintáctico, así como un analizador de charts bidireccional dirigido por islas que usa los mencionados modelos probabilísticos con el fin de guiar el proceso de reconocimiento.
Palabras clave: Análisis sintáctico bidireccional, análisis dirigido por islas, modelado probabilístico del lenguaje

Abstract: The core of this thesis presents and evaluates two methods for stochastically modelling bidirectionality in parsing, as well as a bidirectional island-driven chart parser that uses such stochastic models to guide the recognition process.
Keywords: Bidirectional parsing, island-driven parsing, stochastic language modelling

1 *Summary*

This thesis developes a complete methodology for bidirectional island-driven parsing of natural language. It includes the definition of two methods for stochastically modelling bidirectionality in parsing, as well as a bidirectional island-driven chart parser that uses such stochastic models to guide the recognition process. This bidirectional parser starts the analysis process from certain dynamically determined positions of the sentence (that is, the islands), proceeding then in both directions. Our framework accounts for bidirectional expansion of partial analysis, which improves the predictive capabilities of the system.

Both stochastic models are supervised, starting from a stochastic context-free grammar, and then adding the corresponding stochastic parameters of each model. The models provide for the probability of extension of each island and active edge in the chart structure to both sides. In order to do it, they use information based on either the stochastic grammar (in the case of the local model) or both the grammar and the islands which are immediately adjacent to the one being considered (in the case of the *neighbouring* model).

The stochastic models aim at defining scoring functions (FOMs) to drive the bidirectional parsing algorithm to find the *best-first* parse. The developed chart parser can use such models either independently or in combination. In fact, the structure of the parser has been designed in such a way that it allows for the implementation of both the two unidirectional non-stochastic approaches considered as baselines (bottom-up and top-down) as well as the variety of different bidirectional strategies regarded. The latter include head-driven methodology, island-driven pure methodologies (using only either local or *neighbouring* models), and hybrid methodologies which combine either both models or an stochastic model with a unidirectional approach. The election of the specific hybrid combinations evaluated has been heuristic, since it stems from the necessity of optimising the performance of the *neighbouring* model. It includes back-off from *neighbouring* to local, thresholding of the *neighbouring* parameters, and smoothing of the *neighbouring* probabilities.

Island-driven methodology implies the existence of a method in order to select the initial islands in the sentence being parsed, from which the analysis will proceed. In this thesis, several alternatives have been considered, thouhg only three of them have been completely evaluated in our framework, namely selecting as islands the nonambiguous words

[*] Tesis leída el 19 de Marzo del 2003.

ISSN 1135-5948

in a morphologically analysed but non-tagged sentence, selecting as islands the chunks in a previously chunked sentence, and simply selecting the islands according to their category.

The work in this thesis can be considered as eminently heuristic, since the island-driven methodology presents so many parameters or degrees of freedom, that only through testing and evaluation can one try and find the best alternatives as to the stochastic models, the strategy of the parsing algorithm, and the methodology of selection of the initial islands. Therefore, extensive experimentation has been carried out, and we have been specially concerned with the design of an adaptive environment in which all these parameters can be easily changed.

The system has been trained and tested over two wide-coverage grammars and real corpora: Spanish Lexesp and English Penn Treebank. Performance has been measured in terms of the average number of edges created in order to complete the first parse of the sentence. Parsing performance has been analysed according to several metrics of the input sentence (sentence length, ambiguity rate, island density, etc.). Whatever the island selection methodology, our approaches dramatically outperform both baselines, top-down and (specially) bottom-up strategies.

As mentioned, several hybrids methods which combine local and *neighbouring* approaches have also been defined and evaluated (using the first island-selection strategy, nonambiguous words). Their performance always improves the single ones' (both the baselines and the previous pure stochastic methods). In fact, all the hybrid methods outperform the optimal pure approach, *neighbouring*, by between 46% and 48%. The optimal method is the hybrid *neighbouring* which combines controlled back-off to local with smoothing of the *neighbouring* probabilities. We conclude that, with the first island-selection criterion, the *neighbouring* approach is the optimal, as long as its main drawback, data sparseness, is overcome by smoothing somehow (with any of our proposed methods) the learnt stochastic parameters.

The second island-selection strategy, consisting in preceding the island-driven parser by a chunking process for identifying the initial islands, has been proved even useful for improving parsing performance without loss of coverage, as the chunking process can be carried out quite straightforwardly in a very efficient way. The use of this more informed proposal provides an even more significant improvement on performance. Both the alternatives of using only base noun-phrases as chunks and using also other types have been tested, the former approach obtaining the best results, reducing by 8 the average number of edges with respect to baseline bottom-up.

Besides evaluating performance, all the approaches have been also evaluated as to other quality measures, namely the likelihood of the parses and their accuracy (precision and recall). Local method presents the best results for all the island-selection strategies (these results being optimal for the local model plus the chunking approach).

We have consequently demonstrated that, whatever the island-selection strategy, our island-driven methodology improves the efficiency of the usual unidirectional techniques, and that the percentage of improvement increases with the length of the sentence. Such a contribution is highly relevant, considering that stochastic context-free parsing with large (real-sized) grammars is a problem that might not be tractable for long (real-sized) sentences. Moreover, these natural language sentences might be corrupted (mainly when dealing with speech, but also whenever we might find non-grammatical sentences), which would render the application of unidirectional strategies impossible or, at least, much harder.