IL PARSER BASATO SU REGOLE DEL GRUPPO NLP DELL'UNIVERSITÀ DI TORINO

THE RULE-BASED PARSER OF THE NLP GROUP OF THE UNIVERSITY OF TORINO

LEONARDO LESMO

SOMMARIO/ ***ABSTRACT***

Questo articolo descrive l'architettura generale del parser sviluppato dal gruppo di Elaborazione del linguaggio Naturale del Dipartimento di Informatica dell'Università di Torino. Si tratta di un parser che opera in due fasi principali: una fase di chunking e una fase di aggancio di dipendenti verbali. Le operazioni sono guidate da regole sviluppate manualmente e da informazioni relative alla sottocategorizzazione verbale.

In this paper, we describe the architecture of the parser developed by the NLP group of the Department of Informatics of the University of Torino. It is a parser who carries out the analysis in two main phases: a chunking phase and a phase of attachment of dependents to verbs. The operations are driven by rules manually developed and by data about verbal subcategorization.

Keywords: chunk-based parsing, verbal classes.

1. Introduction

The parser described herein is a wide coverage practical parser, which has been applied to various domains and has been the starting point for the development of TUT (the Turin University Treebank for Italian). The overall architecture is depicted in fig.1. This paper does not cover all phases of the analysis that precede the parsing proper, i.e. the morphological analysis and the POS tagging.

All rules and knowledge bases used by the parser have been developed manually. The various KB's are largely multilingual, since they have been applied, with suitable extensions, to Catalan, English, and Spanish.

2. The chunker

The chunking module scans the input sentences N times, where N is the number of POS that can act as head of substructure. Currently, N=12. The list of POS is ordered according to a "level of complexity", where the POS who can govern simpler structures occur first.

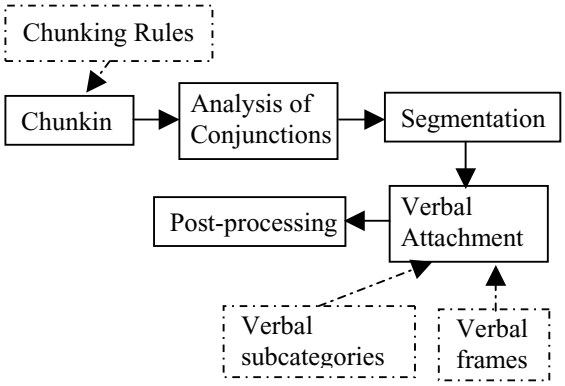

Fig.1 – Architecture of the parser

A "rule packet" is associated with each POS. A packet includes all the rules that describe the possible dependents of a word of that POS. If the i-th element of the list of POS is pos_i , when, during the i-th scan of the sentence, a word w_k of POS pos_i is found, all rules of the packet associated with pos_i are applied.

Basically, a rule can inspect the immediate context of w_k, evaluating, on the surrounding words, the pre-conditions of the rule. If, for a word w_j the preconditions are satisfied, then w_j is linked as a dependent of w_k, labelling the link with the arc name appearing as consequent of the rule. An example rule is:

```
(NOUN common
      (precedes (ADJ qualif T (#\- #\' #\"))
            (ADJ ((type qualif)
                  (agree)))
      ADJC+QUALIF-RMOD))
```

This rule specifies that if a potential Head which is a common NOUN is preceded by a qualificative adjective that syntactically agrees in gender and number (possibly with zero or more qualif adj and some punctuation marks in between), then the adjective must be linked to the noun with the label ADJC+QUALIF-RMOD.

3. Analysis of conjunctions

After the previous step, various chunks have been identified in the sentence. Now, the parser looks for conjunctions. For instance, in "Il ragazzo e la ragazza si incontrarono al bar" [the boy and the girl met at the bar], the result of this step is to build up a single group, including the two chunks "il ragazzo" and "la ragazza". In this way, a single dependent is associated with the verb.

It is not possible here to describe in detail the complex (and error-prone) process that handles conjunctions. In short, it first looks for the possible second conjuncts (after the conjunction), on the basis of the chunks. Then, for each of them, inspects the previous part of the sentence in search for a compatible first conjunct. Note that the elements that can be conjoined vary considerably (as "She and the horse", "rapidly, but with great care") and the two conjuncts can be separated by a large portion of text. The match is based on a set of (procedural) heuristic rules.

After all possible pairs of first and second conjunct have been found, a set of preference rules is applied in order to choose the best solution.

4. Segmentation

The segmentation process detects the set of dependents for each verb occurring in the sentence. This is made procedurally by moving back and forth from each verb.

However, each verb is handled without taking into account mutual influences. Since the verbs are inspected from left to right, most of the processing is made by looking for the dependents that follow the verb. In this search, all unlinked material is collected (chunk heads), until some "barrier" is found. Examples of barriers are interrogative adverbs, subordinating conjunctions, etc.

5. Verbal attachment

After all candidate dependents for a verb have been found, the arc labels have to be determined. This is made by using two knowledge sources: the first one associates with each verb one or more names of subcategorization classes. The second one specifies the class features.

An example of class specification is the following:

Affermare: refl full-basic-trans trans-dir-disc

Here, it is stated that the verb "affermare" belongs to three classes, i.e. *refl* (reflexive: 'have success'), *full-basic-trans* (basic transitives admitting sentential objects: 'to state'), *trans-dir-disc* (transitives admitting direct discourse, i.e "he stated 'you are very bright'").

The definition of the classes is given in terms of a hierarchy, where a special inheritance mechanism is defined (the hierarchy does not encode a subclass (is-a) relation). The root of the hierarchy is the *verbs* class, with no specific info. The subclasses of *verbs* are: *no-subj-verbs*, *subj-verbs*, *obj-verbs*. The first of them refers to verbs that do not take a subject (as, in Italian, 'piove' - to rain - and 'bisogna' – must). *subj-verbs* include the definition of the possible 'basic' realizations of subjects, in terms of their heads. They include (bare) nouns, pronouns, articles (which govern the associated noun), etc. It is also stated that the subject must agree with the verb. Similarly for *obj-verbs* (definition of direct object). With respect to standard subcategories, *intransitives* are equivalent to *subj-verbs*, while *transitives* are defined as the daughter of *subj-verbs* and *obj-verbs*, so that they 'inherit' the presence of a subject and of a direct object.

On the basis of the class definition and of the definition of a set of *transformations* (e.g. passivization) a set of possible *surface realizations* is automatically generated. These are finally matched against the actual dependents hypothesized for a given verb (according to its class) in order to find out the arc labels. The classes only encode the complements of verbs, so that the match must also take care of the possible presence of *adjuncts*.

6. Post-processing

Because part of the process is based on heuristic rules, it may happen that, at the end, some elements remain unattached. In this last step, some simple heuristic rules are applied to determine the most reasonable governor for these elements, so that the final structure is fully connected to the root of the tree. A final check, that has not yet been implemented, should take care of detecting cycles and non-projective substructures. More details about the parser appear in the references below.

REFERENCES

[1] L. Lesmo and V. Lombardo: "Transformed Sub-categorization Frames in Chunk Parsing", Proc. 3rd Int. Conf. on Language Resources and Evaluation (LREC 2002), Las Palmas, 2002, 512-519.

[2] L. Lesmo, V. Lombardo & C. Bosco: "Treebank Development: the TUT Approach", in R.Sangal and S.M.Bendre (eds.): Recent Advances in Natural Language Processing, Vikas Publ. House, New Delhi, 2002, 61-70.

CONTACT

LEONARDO LESMO
Dipartimento di Informatica, Università di Torino
Corso Svizzera, 185
10149 Torino
Email: lesmo@di.unito.it

LEONARDO LESMO is Professor of Man-Machine Interaction. He works on NLP, Agent-Based Models of Communication and Legal Ontologies. He is Vice-President of the Center for Cognitive Science, member of the Steering Committee of AI*IA and of the Italian Association of Cognitive Sciences. He has been local coordinator of various national and international projects.