IL POS TAGGER DEL GRUPPO NLP DELL'UNIVERSITÀ DI TORINO

THE POS TAGGER OF THE NLP GROUP OF THE UNIVERSITY OF TORINO

LEONARDO LESMO

SOMMARIO/ ***ABSTRACT***

Questo articolo descrive l'architettura generale del POS tagger sviluppato dal gruppo di Elaborazione del linguaggio Naturale del Dipartimento di Informatica dell'Università di Torino. Nel testo viene inoltre descritto l'analizzatore morfologico e il trattamento delle multi-word. Il tagger è basato su regole sviluppate manualmente ed è stato applicato, con opportune variazioni, anche a lingue diverse dall'Italiano.

This paper describes the architecture of the POS tagger developed by the NLP Group of the Dipartimento di Informatica of the University of Torino. A description of the morphological analyser and of the way multi-words are handled is also included. The tagger is based on hand-written disambiguation rules and has been applied, with suitable adaptations, also to languages other than Italian.

Keywords: POS tagging, morphological analysis

1. Introduction

The POS tagger that is described herein is a rule-based system, that takes as input the result of the morphological analysis of a sentence. This result may include multiple entries for each word, when an ambiguity is present. The output of the tagger is a sequence of single entries, each of which is associated with an input word.

The term "Part of Speech" is only partially correct, since the tagger takes care not only of the choice of the most reasonable POS label, but also of the multiple readings possibly arising from intra-category ambiguities, as in case of "sono", which is a verb, but can be an auxiliary or a main verb, and corresponds both to the 3rd person plural form (they are) and to the 1st person singular form (I am).

Multi-words are accounted for in a special way, since the tagger always prefer a multi-word to its individual counterparts.

2. Morphological Analysis

This first phase is preceded by the *tokenization* of the input text. This aims at identifying *tokens* in the input. The tokens can be of different types, e.g. GW (general word), PN (proper name), DATE, NUM, etc. The type assignment is not deterministic; for instance, capitalized words get both the GW and the PN type. These ambiguities are solved by the POS tagger. During tokenization, sentence boundaries are hypothesized, since all subsequent steps (tagging and parsing) work on individual sentences.

All tokens of GW type undergo the morphological analysis. The dictionary is based on word stems: each entry includes data about the features of the associated lemma(ta) and about the *morphological class*. An example of a dictionary entry is:

(ris ((riso cat noun classe (2))
(ridere cat verb classe (8 (c (1 3 6) i)) transitive no)))

Here, the stem is *ris*, with which two lemmata are associated: the noun "riso" (rice or laugh) and the verb "ridere" (to laugh). "riso" has the morphological class 2, which states that the possible endings are –o (for masculine singular) and –i (for masculine plural). The class codes for the verb are more complex, since they state that the stem *ris* applies only to some forms of the "passato remoto" tense (*c*) and to the past participle (*i*).

The definition of the morphological classes appears in a suffix table; the morphological analysis inspects the word token starting from the end, and extracts all possible suffixes. Then, the stems are matched against the dictionary. The task is made a bit more complex by the possible presence of enclitics; in Italian, they can be attached to the end of verbs (e.g. *prendi-me-lo*: take – for me - it).

3. Multi-words

To Multi-words are represented as an automaton of word forms. The POS tagger module, before applying the

standard rules described in the next paragraph, checks if, at the beginning of the current portion of sentence there is a sequence of forms that can be recognized by the automaton. Since the tagging proceeds from left to right, the beginning of the "current portion" is moved across the sentence, and all possible multi-words are detected. A possible limitation of the tagger is that multi-words are always preferred, so that, in case of ambiguity, a 'standard' reading of the form sequence is never selected.

4. POS Rules

The POS tagging rules are associated with specific ambiguity sets. By "ambiguity set", we mean the different possible categories and/or features of the interpretations of a word. For instance, the word "rosa" is associated with three POS:

a. noun (the flower)
b. adj (the colour)
c. verb (past participle of the verb "rodere")

On the basis of the result of the lexical access, the "ambiguity set" is extracted. In the example above, the ambiguity set is *(adj noun verb)*.

For each ambiguity set, a packet of rules is defined. For example, the packet for (*adj noun verb*) includes 25 rules.

An example of a rule is reported below:

```
(adj-noun-verbr11
  :if '(and (prevcat 'adv)
            (prev2type 'aux)
            (currmood '(participle gerund)))
  :then 'verb
  :CF 'U))
```

In this rule, three predicates are involved:

- *prevcat*: it checks if the POS of the preceding word is the one given (in this case "adv")
- *prev2type*: it checks if the syntactic subtype of the word before the preceding word (back of 2 words) is the one given (in this case "aux"; since "aux" is value applying only to verbs, the involved word must be a verb)
- *currmood*: the mood of the current word must be "participle" or "gerund". Again, this implicitly refers to the verbal interpretation of the current word, since adjectives and nouns do not have mood

If the predicates are satisfied, then the assigned category is *verb*. A certainty factor is also included (U = uncertain), but its role is limited, since the rules are mainly manually ordered (three values are defined: C - certain, A - almost certain, U - uncertain).

Currently, there are 56 different predicates involved in the rules, enabling them to check various features of the two words preceding and the two words following the word to be disambiguated. Among them, 5 predicates enable the tagger to have a larger window on the sentence; in particular, they check if the sentence possibly is interrogative (has a final question mark) and if there are specific types of verbs inside it.

It must be observed that the tagging rules are applied left-to-right. This means that, when the POS of a word is chosen, the 2 words preceding the one under analysis have already been disambiguated, while the two words after it are still ambiguous.

Finally, we note that the final aim of the POS tagger is to select a single word, so that part of the tagging rules are devoted to inter-categorial ambiguities (as, for instance, "sono" which is the 1st person singular present and the 3rd person plural present of the verb "essere" - to be). In these cases, the rules apply to different features, as the syntactic gender or number, the tense for verbs, or the distinction between common nouns and proper names.

5. The involved knowledge bases

The standard dictionary includes 23.400 roots, corresponding to about 25.000 lemmata. A dictionary of proper names is also used, which includes 570 names (first names of persons, artists, and geographical places).

Multi word expressions are encoded as automata of words. 305 such MWE are defined, plus 110 proper names MWE referring to places (as "Los Angeles") or artists (as "William Shakespeare"). The POS tagging rules are 614, grouped in 90 packets. 24 of these packets refer to inter-categorial ambiguities.

Some semantic-biased word classes are also defined. They include some words referring to places (e.g. street, garden, …), times (e.g. Monday, august, …). These classes are limited to some specific set of words, but they could be seen as a first step towards the exploitation of semantic infos during the tagging phase.

The reference below gives some more details about the tagger, and describes some experiments that were made to apply automatic tools to rule refinement.

REFERENCES

[1] G.Boella, L.Lesmo: "Automatic Refinement of Linguistic Rules for Tagging", Proc. 1st Int.l Conf. on Language Resources and Evaluation, Granada, 1998, 923-929.

CONTACT

LEONARDO LESMO
Dipartimento di Informatica Università di 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.