Rules

for the optimisation of the automatic inflexion

of Italian "co" and "go" nouns and adjectives

Using the web as a linguistic resource

Lucia SANTORUM
Université Paris 4 - Sorbonne
28, rue Serpente
75006 Paris
luciasantorum@gmail.com

Gregory GREFENSTETTE
CEA-LIST/LIC2M
92265 Fontenay aux Roses Cedex
grefenstetteg@cea.fr

Keywords - Mots-clés

Italian, inflexion, automatic inflexion, optimization, cooccurrence, rule, exception, plural, consonant, vowel, affix.
Italien, flexion, flexion automatique, optimisation, cooccurrence, règle, exception, pluriel, consonne, voyelle, affixe.

Abstract – Résumé

This paper presents experimental results for optimizing the inflexion rules of some Italian nouns by reducing the number of exceptions governing their automatic inflexion. The central result of this paper is that the inflexional behaviour of Italian nouns ending in "co" and "go" can be related to the letters immediately preceding the "c" or "g" avoiding the need for phonetic or other considerations which are difficult to implement in an automated system for natural language processing.

Les résultats présentés dans cet article proposent une nouvelle règle contribuant à l'optimisation de la flexion automatique de l'italien. Plus précisément, nos observations montrent la possibilité de réduire le nombre d'exceptions pour une certaine catégorie de mots, avec, pour conséquence, la réduction de la taille des fichiers et donc de leur temps de traitement. Notre approche se fonde sur l'observation du comportement des mots terminant par "co" et "go" en fonction uniquement des graphèmes qui les précèdent et non de considérations à caractère phonologique ou sémantique.

Overview of the question

Systems which create lexical resources for analysing natural language texts must possess a morphological analysis tool that converts input strings found in text into lemmatized form used by the system itself. This conversion can be done via a finite-state automata that implicitly contains rules converting inflected forms into lemmatized forms (Ferri et al, 1997), or it can be done by comparing surface, or by a programmed implementation of explicit rewriting rules (Johnson, 1972) or by producing a full form lexicon that explicitly maps surface forms to lemmatized forms (Ide and Veronis, 1994). In all these cases, rules are needed that map from surface form phonological elements (or strings of characters) into other elements, those found in the lemmatized form. The question addressed in this paper is the definition of the inflexion rules followed by Italian masculine nouns and adjectives with the affix "o" in the case the affix is preceded by the letters "c" and "g". To our knowledge operational rules for performing this inflexion for this large class of words has never been described in a way amenable to computer programming. Our presentation is divided into three parts: in part one, we explain the traditional rule proposed in standard Italian grammars to govern these morpheme combination and we illustrate its limitation for NLP. We then present an initial solution to this problem, using the Web to make the choice as to the proper inflexions. In part three, we present the effective rule that we have discovered that solves the problem in a clean and efficient way for NLP systems.