CORISTAGGER: UN POS TAGGER PERFORMANTE PER L'ITALIANO

CORISTAGGER: A HIGH-PERFORMANCE POS TAGGER FOR ITALIAN

FABIO TAMBURINI

SOMMARIO/*ABSTRACT*

Questo contributo presenta una versione evoluta di CORISTagger [1], un *PoS-tagger* per la lingua italiana. Il sistema è composto da un annotatore basato su modelli di Markov i cui risultati vegnono rielaborati da un *Transformation Based tagger*. L'uso di questa combinazione di tagger in congiunzione con un potente analizzatore morfologico ha permesso al CORISTagger di ottenere ottime prestazioni nel *PoS-tagging task* di EVALITA 2007.

This paper presents an evolution of CORISTagger [1], an high-performance PoS-tagger for Italian. The system is composed of an Hidden Markov Model tagger followed by a Transfomation Based tagger. The use of such a stacked structure paired with a powerful morphological analyser, allowed the tagger to obtain very good performances in the EVALITA 2007 PoS-tagging task.

Keywords: PoS tagging, HMM tagger, Rule-based tagger.

1 Introduction

The tagger presented in this paper is an evolution of the tool developed inside the CORIS project [1]. The earlier version of this tagger were based on a single HMM system, but for this task a more complex tagging structure has been developed (see next section).

During the development phase, the Development Set (DS) has been split into two parts respecting the same proportion between DS and the Test Set (TS) described in the task Guidelines. Then the whole system has been trained and tested and various improvements, regarding both the system structure and the single system components, were introduced and carefully checked.

During the final evaluation, the system was trained using only the development set, no other textual resources have been used for this evaluation campaign.

2 Overall Tagger Structure

The overall tagger structure is depicted in figure 1. The whole tagger consists of two different tagging models stacked in order to achieve better performance. A standard second order HMM tagger [1], enriched with numerous smoothing techniques, produces a first-step output that feeds a transformation-based tagger (fnTBL [2]). The idea is to use the rule-based tagger to correct the mistakes done by the first step HMM tagger. By learning only the appropriate set of rules to correct the first step errors, this second part can benefit of an enlarged context horizon. Moreover the training phase can be pushed forward to a level unreachable with a single rule-based tagger starting from a preliminary tagged corpus annotated with the most frequent tag, as in the standard use of such models.

Both taggers can benefit from the use of a morphological analyser based on a huge lexicon.

2.1 The Morphological Analyser

The whole system uses a large lexical resource embodied into a powerful morphological analyser. The underlying model is the TFS-formalism; a huge lexicon composed of about 120,000 lemmas, slightly smaller that the De Mauro-Paravia online dictionary, has been created and it is used in every phase of the disambiguation process.

As showed in [1], the use of such a huge lexical resource reduces the number of unknown words essentially to proper names (78%), common nouns (10%) and adjectives (7%). Thus, when the tagger has to process a word not recognised by the morphological analyser, we can apply simple heuristics to guess the available PoS tags for this token. If the first character is uppercase and the token is not at the beginning of a sentence, then the tagger assigns to it the tag corresponding to proper names, else both tag for nouns and adjectives are assigned and the disambiguation task is left to the Viterbi algorithm. The heuristic is very simple, but, due to the large lexical resource used, we reach good performances, as we can see in section 3.

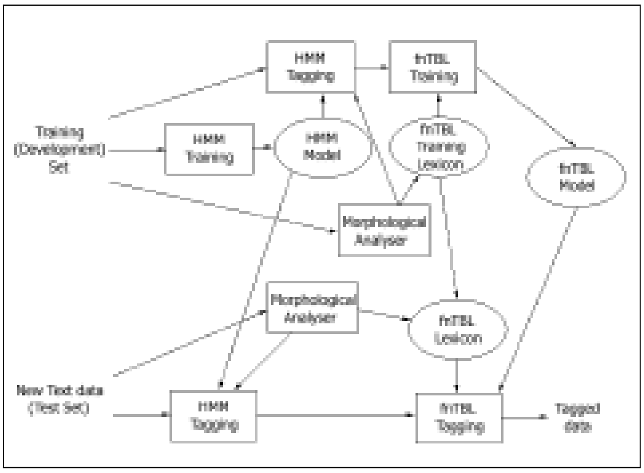

Figure 1: The overall tagger structure.

2.2 The HMM tagger

The core part of the *CORISTagger* is composed of a standard second-order HMM tagger. Various smoothing techniques were applied in order to avoid the classical problems of such methods, in particular the underflow problem, while the data sparseness was corrected applying the techniques suggested for example in [3] and [4] (a log scale transformation of the governing equations and the interpolation of n-gram frequencies).

2.3 fnTBL tagger

fnTBL is an open-source package implementing a machine learning technique called transformation based learning (TBL), first introduced by Eric Brill in 1992. It is mainly based on the idea of successively transforming the data in input to correct the error that gives the biggest error rate reduction. The transformation rules obtained are usually few and meaningful.

fnTBL allows for a number of special configuration options that make it ideal for our purposes. It requires an input file already tagged with the most frequent tag, then it was very easy to stack it after the HMM tagger and instruct it to use the HMM tagger output instead the most frequent tag. Moreover, it allows for an easy configuration of the context features considered for the tagging task. We maintained the rule templates proposed by the standard package, but we made a longer training phase, so that the system learnt rules that corrected at least 2 errors.

3 Performances and Discussion

Table 1 shows the evaluation results for *CORISTagger* with respect to the two evaluation metrics. The performances are very high, both as absolute value when compared to the state-of-the-art tagging results for English and when compared to the other participants of EVALITA 2007 campaign. The presented PoS-tagger ranked 4th for EAGLES-like tagset and 3rd for the DISTRIB tagset.

Stacking fnTBL over the HMM tagger improved the overall system performances giving a reduction in error rate on Tagging Accuracy larger than 0.5%.

Table 1: *CORISTagger* (*UniBoDSLO_Tamburini_POS*) results with respect to Tagging Accuracy (TA) and Unknown Words Tagging Accuracy (UWTA).

Tagset	**TA**	**UWTA**
EAGLES-Like	97.59	92.16
DISTRIB	97.31	92.99

REFERENCES

[1] F. Tamburini. Annotazione grammaticale e lemmatizzazione di corpora in italiano. In R. Rossini, *Linguistica e informatica: multimedialità, corpora e percorsi di apprendimento.* pp. 57-73, Bulzoni, Roma, 2000.

[2] G. Ngai, and R. Florian. Transformation-based learning in the fast lane. *In Proc. of NAACL-2001*, pp. 40-47, 2001.

[3] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun. A Practical Part-of-Speech Tagger. *In Proc. ANLP'92*, pp. 133-140, 1992.

[4] T. Brants. TnT – A Statistical Part-of-Speech Tagger. *In Proc. ANLP 2000*, pp. 224-231, 2000.

CONTACT

FABIO TAMBURINI
DSLO - University of Bologna
Via Zamboni, 33, I-40126, Bologna
Email: fabio.tamburini@unibo.it

FABIO TAMBURINI is assistant professor at DSLO, University of Bologna, Italy. His main interests are in computational linguistics, speech processing, and corpus linguistics.