PoS-tagging Italian texts with *CORISTagger*

Fabio Tamburini

DSLO, University of Bologna, Italy
fabio.tamburini@unibo.it

Abstract. This paper presents an evolution of *CORISTagger* [1], an high-performance PoS-tagger for Italian developed at the University of Bologna. The system is composed of a second-order Hidden Markov Model tagger followed by a Transformation Based tagger. The use of such a stacked structure, paired with a powerful morphological analyser based on a large lexicon composed of 120,000 lemmas, allowed the tagger to obtain good performances in the EVALITA 2009 PoS-tagging task. The performances of the tagger and the most common classification errors are discussed in detail.

Keywords: PoS-tagging, HMM tagger, transformation-based tagger, morphological analyser.

1 Introduction

The tagger presented in this paper is an evolution of the tool developed inside the CORIS project [1]. It has been successfully used to tag the CORIS/CODIS corpus [2], as well as in lots of other projects, and it participates also to the EVALITA 2007 campaign [3] obtaining very good results. The earlier versions of this tagger were based on a single Hidden Markov Model system, but for the pos-tagging task in the EVALITA campaigns a more complex tagging structure has been developed, adding a Transformation Based tagger after the HMM tagger in a stacked structure [8].

During the development phase, 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.

For the final evaluation, we participated only to the *open-task* producing two different *runs*: in the first (Tamburini1) the system was trained using both the training set and development set provided by the organisation, while in the second (Tamburini2) we added 64876 tokens extracted from the texts used for the EVALITA 2007 pos-tagging task, after having converted the tagset using a semi-automatic procedure. Because of the differences in granularity between the tagsets used in the last two EVALITA campaigns, the preparation of the texts for this second run required a careful revision of the results produced by the automatic conversion procedure used to convert the EVALITA 2007 *EAGLES-like* tagset to the EVALITA 2009 *TANL* tagset.

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. This solution has been successfully experimented during the EVALITA 2007 campaign.

A standard second order HMM tagger [1], enriched with numerous smoothing techniques, produces a first-step output that feeds a transformation-based tagger (fnTBL [4]). 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 to carefully handle the words not belonging to the very small training set provided by the organisation composed of about 118.000 tokens.

2.1 The Morphological Analyser

The whole system uses a large lexical resource embodied into a powerful morphological analyser. The underlying model is the Typed-Feature-Structure (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. With regard to open-class words, it contains about 11,000 verbs, 30,000 adjectives, 70,000 nouns and 5,000 adverbs.

The morphological analyser is able to provide a complex set of information for each analysed word: pos-tag, lemma, mood, time, person, gender, number, etc. are only some of the information available using such tool.

As showed in [1], the use of such a huge lexical resource allows a coverage of more than 98% of text tokens, and it 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 the tag for nouns, adjectives and foreign words are assigned and the disambiguation task is left to the stacked taggers. The heuristic is very simple, but, due to the large lexical resource used, we can usually reach good performances. Considering the morphologically rich nature of the TANL tagset, the application of such heuristic would require to establish also the number and gender of the hypothesised noun or adjective. Such information is unfortunately not available for unknown words and no further simple heuristics can be defined without increasing the number of hypothesised tags. For this reason we decided to assign gender *male* and number *singular* to every noun or adjective hypothesised by the heuristic used to process unknown words.

The need to process a fine-grained tagset enriched with lots of morphological information forced us to revise the morphological analyser and add, in some cases, all

such information, obtaing a more complete and reliable resource useful in various NLP tasks.

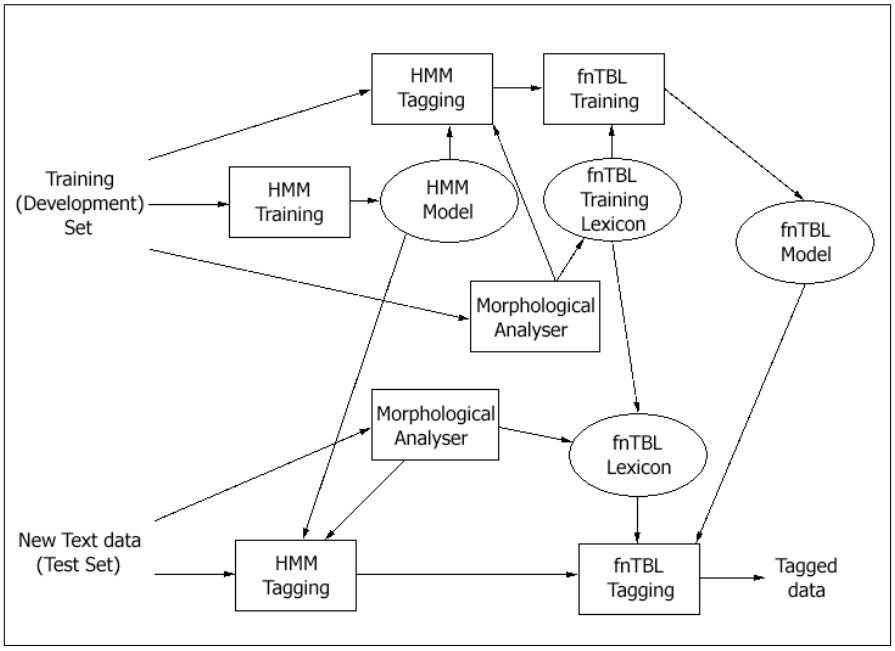

Fig. 1. The overall tagger structure.

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 and the data sparseness were corrected applying the techniques suggested, for example, in [5] and [6] (a log scale transformation of the governing equations and the interpolation of n-gram frequencies).

2.3 fnTBL tagger

fnTBL [4] 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, containing templates spanning form -3 to 3 positions around the examined word, but we made a longer training phase, so that the system learnt rules that corrected at least 2 errors.

3 Results and Discussion

Table 1 shows the evaluation results for *CORISTagger* with respect to the evaluation metrics. The performances are quite high when compared to the state-of-the-art tagging results for Italian considering the very rich tagset used for this task.

Analysis of results. The overall picture of *CORISTagger* results when compared to the other participants of EVALITA 2009 campaign on the morphed (POS) tagset (see [7] for the description of the pos-tagging task results) deserves some further discussion. All the systems results are clearly divided into two groups: the first group of systems exhibit an accuracy results around 96.5%, while the second group around 96%. *CORISTagger* is part of this second group of systems. The differences inside the two groups, when examining the number of misclassified tokens, is very limited, and spans around 10/15 tokens.

The test data-set was composed by only 5066 tokens, too few, in our opinion, for producing results able to perform an in depth evaluation of the differences between the participating systems.

The performances obtained by *CORISTagger* using the coarse-grained tagset are quite better, allowing it to place in the middle of the ranking list.

Table 1. *CORISTagger* results with respect to fine-grained morphed tagset (POS) and the coarse-grained tagset (CPOS) at EVALITA2009 open task.

OPEN TASK	**POS accuracy**		**CPOS accuracy**		**Unkn. POS acc.**		**Unkn. CPOS acc.**	
Tamburini 1	95,93%	4719/4919	96,40%	4742/4919	90,95%	794/873	92,67%	809/873
Tamburini 2	95,63%	4704/4919	96,16%	4730/4919	91,07%	795/873	92,78%	810/873

Analysis of errors. Table 2 outline the most common errors performed by the *CORISTagger* considering both the submitted runs. Let us have a closer look at the errors for the *RUN1*, the most effective one:

- the most common error involves a confusion between proper names and common nouns; this is probably due to the simple heuristic we used to handle unknown words not recognised by the morphological analyser.

– the second type of errors regards classical problematic categories for pos-taggers: adjectives *vs* past participles (quite complex also from a theoretical point of view) and adjectives *vs* common nouns.
– the third, more problematic, group of errors regards the misclassification of some relative pronouns, adverbs and prepositions. Examining also the errors done by the system at the *RUN2* we can note that this problem is even worse becoming the most common kind of error done by the tagger. The reason for this kind of misclassification can be brought back to the classification of lemmas in the morphological analyser. Methodological, or theoretical, differences in the attribution of lexical classes lead to different results when the tagset for which the system is designed has to be changes to adapt it to a new situation. A correct, and satisfactory, mapping between the two tagsets is often not sufficient to adapt the system, and a complete, very long, revision of the external resources for adapting them to the different theoretical view of the new task is usually required. Unfortunately, confining our discussion to the EVALITA evaluations, such theoretical claims are not usually expressed clearly, but they are left implicitly described into the training data, making the process of designing a successful and complete porting of the system more difficult.

Table 2. *CORISTagger* most common classification errors for both runs submitted for the evaluation at the open task.

Run 1			**Run 2**		
11	SP	Sms	13	B	E
10	Ams	Sms	11	SP	Sms
9	Ams	Vpsms	10	Ams	Vpsms
7	SP	Sfs	10	Ams	Sms
7	CS	PRnn	9	SP	Sfs
7	Amp	Smp	6	VAip3p	Vip3p
7	Afp	Vpsfp	6	Amp	Vpsmp
6	B	E	6	Amp	Vpsmp
6	Amp	Vpsmp	5	CS	PRnn
5	VAip3p	Vip3p	5	B	CC
5	Ans	Sms	5	Ans	Sms
5	Afs	Vpsfs	5	Amp	Smp
			5	Amp	Anp
			5	Afp	Vpsfp

4 Conclusions

This paper presented *CORISTagger* a PoS-tagger specifically developed and tailored for the Italian language and its performance results at the EVALITA 2009 evaluation campaign.

The results and the classification errors have been discussed and, although the performances are not in top rank of the classification list, the differences with the other

systems participating to the evaluation were negligible, especially considering the small amount of data available for this challenge.

Further developments regard the introduction of more reliable heuristics to handle unknown words and the testing of different stacking and voting schemes in the spirit of the work done in [8].

References

1. Tamburini, F.: Annotazione grammaticale e lemmatizzazione di corpora in italiano. In: Rossini, R. (ed.), Linguistica e informatica: multimedialità, corpora e percorsi di apprendimento, pp. 57-73, Bulzoni, Roma (2000)
2. Rossini Favretti, R., Tamburini, F., De Santis, C.: CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model. In: Wilson, A., Rayson, P., McEnery, T. (eds.), A Rainbow of Corpora: Corpus Linguistics and the Languages of the World, pp. 27-38, Lincom-Europa, Munich (2002)
3. Tamburini, F.: EVALITA 2007: the Part-of-Speech Tagging Task. In: Intelligenza Artificiale, vol. IV(2), pp. 57-73 (2007)
4. Ngai, G., Florian, R.: Transformation-based learning in the fast lane. In: Proceedings of 2nd Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2001), pp. 40-47, Pittsburgh, PA (2001)
5. Cutting, D., Kupiec, J., Pedersen, J., Sibun, P.: A Practical Part-of-Speech Tagger. In: Proceedings of 3rd Applied Natural Language Processing Conference (ANLP'92), pp. 133-140 (1992)
6. Brants, T.: TnT – A Statistical Part-of-Speech Tagger. In: Proceedings of 6th Applied Natural Language Processing Conference (ANLP-2000), pp. 224-231 (2000)
7. Attardi, G., Simi, M.: Overview of the EVALITA-2009 PoS Tagging Task. Same volume, (2009)
8. van Halteren, H., Zavrel, J., Daelemans, W.: Improving Accuracy in Word Class Tagging through the Combination of Machine Learning Systems. In: Computational Linguistics, 27, pp. 199-229 (2001)