# A multistage PoS-tagger at the EVALITA 2009 PoS-tagging Task

Roberto Zanoli and Emanuele Pianta

FBK-irst,
via Sommarive 18, I-38123 Povo (TN), Italy
{zanoli, pianta}@fbk.eu

**Abstract.** This paper presents an experimental system architecture for Part-Of-Speech Tagging for the Italian language, able to manage a large tagset to provide both lexical and morphological information. The tagger was built as a cascade of four classifiers where each classifier in the cascade accepts data from an initial input or the guesses of the previous one, executes its annotation, and sends the resulting data to the next stage, or to the output of the cascade. At the EVALITA 2009 PoS-tagging task the combined classifier attained an accuracy of 96.06% on the open task and of 93.54% on the closed one.

**Keywords:** Part-of-Speech tagging, word-category disambiguation, Italian PoS tagger.

## 1 Introduction

Part-Of-Speech Tagging is the problem of marking up the words in a text as corresponding to a particular part of speech. The most frequently applied techniques for this task are based on Hidden Markov Models [2], and Support Vector Machines (SVMs) [3]. YamCha[1], an open source text chunker for a number of Natural Language Processing (NLP) tasks, was used to implement the system performing the best at Evalita 2007 [5]; drawing on that work, we propose an experimental system architecture to build a multistage PoS-tagger able to manage a large tagset providing both lexical and morphological information. The system was trained on the Evalita 2009 Training Corpus, a corpus annotated using the version of the Tanl tagset [1] that includes morphological features and consists of 328 tags, from 14 basic categories. In the rest of the paper we give further details on the task, the system architecture we used, the feature space, and the results we obtained.

## 2 The task description

In the The EVALITA 2009 Part-Of-Speech Tagging task, systems are required to assign a tag, consisting of a combination of lexical categories (PoS tag) and morphological features to each token. The provided data set consists of articles from the on-line edition of

[1] http://chasen.org/ taku/software/yamcha/

the newspaper La Repubblica[2] and consists in 108,874 word forms, annotated using the Tanl tagset. Tanl provides three levels of POS tags: coarse-grain (14 categories), fine-grain (36 categories), and the morphed tags (328 categories). The evaluation was based on two data sets provided by the organizers: the Training Corpus (TrC) for training systems, and the Test Set (TeS) for the evaluation, whereas a Development Set (DeS) was given for tuning the system. There were two subtasks: a closed task, where participants are not allowed to use any external resources besides the supplied TrC and TeS, and an open task, where participants can use external resources. The systems were evaluated in terms of Tagging Accuracy (it is defined as the percentage of correctly tagged tokens with respect to the total number of tokens) and in terms of Unknown Words Tagging Accuracy (it is defined as the Tagging Accuracy restricting the computation to words present in TeS but not in TrC).

## 3 The system architecture

The tagger was built as a cascade of four classifiers in which the subsequent classifier in the cascade accepts data from an initial input or the guesses of the previous one by means of a guess feature, executes its annotation, and sends the resulting data to the next stage, or to the output of the cascade. Each classifier, based on YamCha to implement the Support Vector Machine, was ordered in terms of increasing specificity such that early classifiers are simple and general, whereas later ones are more specific: a first classifier to recognize the coarse-grain tags, a second one for the fine-grain tags, another one for morphed tags and finally a specific classifier to recognise clitics.

## 4 Experiments and Results

The system was trained on the TrC and configured on the DeS; for this purpose only one experiment was done. The resulting configuration was then tested on the TeS. Concerning the features space, for each word, a rich set of features were extracted: the word itself (both unchanged and lower-cased), prefixes and suffixes (1, 2, 3 or 4 characters at the start/end of the word), prefixes and affixes of the consonant/vowel pattern (1, 2, 3 or 4 items at the start/end of the word, e.g. C CV CVC CVCC, etc.) and orthographic information (e.g. capitalization, hyphenation) for the closed task; in addition, for the open task, the morphological features produced by MorphoPro [4] were added.

**Table 1.** Classifier results on the two subtasks with respect to Tagging Accuracy (TA) and Unknown Words Tagging Accuracy (UWTA). TnT-Tagger was used as a baseline.

| | Closed | | Open | |
|---|---|---|---|---|
| Corpus | TA | UWTA | TA | UWTA |
| DeS(baseline) | 92.96 | 74.02 | - | - |
| DeS | 93.50 | 83.45 | 96.20 | 91.61 |
| TeS | 93.54 | 85.45 | 96.06 | 92.21 |

[2] http://www.repubblica.it/

Depending on the classifier in the cascade, each of these features were extracted for the current word and for the previous and following word; the tag of the two words preceding the current word was considered too. YamCha was then set to work with the PKI algorithm with a first degree of polynomial kernel and pair-wise as method for solving multi-class problems; PKI performs the same results as the original SVMs, whereas PKE, the other algorithm availbale with YamCha, is an approximation of SVMs but is much faster than PKI (3-30 faster). The same system configuration was used for both subtasks, all available features were used without any feature reduction, and no specific method was applied to classify unknown words; prefixes and suffixes for the closed task, and morphological features for the open one, are supposed to be useful for classifying both known and unknown words. Finally the TnT-Tagger was used as a baseline for comparing our results on the DeS.

## 5 Discussion

Comparing the results obtained on the two subtasks, a significant improvement in accuracy was observed moving from the closed to the open task; this has been achieved through the exploitation of morphological information. Morphology also plays an important role in decreasing the difference between the Tagging Accuracy and the Unknown Words Tagging Accuracy: 8.09 for the closed task, and 3.85 for the open one. The comparison of the results of our system on the DeS (93.50) with the results obtained with the TnT-Tagger (92.96) seems to confirm that the system architecture we proposed can deal well with large tagsets, decreasing the effects of data sparseness. The system resulted in middle rank in comparison to the other participants systems, but differently from the closed task in which there was a significant difference in accuracy amongs systems, results on the open task were very close to each other (1.12 the difference in accuracy between the system at the top and at the bottom of the rank).

## 6 Conclusion

We presented an approach to PoS-tagging for Italian that uses a cascade of classifiers to manage large tagsets. A number of features were used without any feature reduction: prefixes, suffixes, and morphological features were considered for classifying both known and unknown words. The system resulted in middle rank in comparison to the other Evalita participants, and for the next future, also considering that only one experiment was done on the DeS, we are going to try optimizing the configuration of each classifier in the cascade and try using a different algorithm for sequences tagging, such as Conditional Random Fields (CRF).

## References

1. Attardi, G., et al.: Tanl (Text Analytics and Natural Language processing). Project Analisi di Testi per il Semantic Web e il Question Answering, `http://medialab.di.unipi.it/wiki/index.php/Analisi_di_testi_per_il_Semantic_Web_e_il_Question_Answering` (2008)

2. Brants, T.: TnT - A Statistical Part-of-Speech Tagger. In: Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000. Seattle, WA (2000)
3. Nakagawa, T., Kudoh, T., Matsumoto, Y.: Unknown word guessing and Part-of-Speech tagging using support vector machines. In: Proceedings of the 6th Natural Language Processing Pacific Rim Symposium, pp. 325-331 (2001)
4. Pianta, E., Girardi, C., Zanoli, R.: The TextPro tool suite. In: Proceedings of LREC, 6th edition of the Language Resources and Evaluation Conference. Marrakech, Morocco (2008)
5. Pianta, E., Zanoli, R.: TagPro: A system for Italian PoS tagging based on SVM. Intelligenza Artificiale, Special Issue on NLP Tools for Italian, vol. IV, issue 2 (2007)