# EVALITA 2007: VALUTAZIONE DI SISTEMI PER L'ANNOTAZIONE DELLE PARTI DEL DISCORSO

# *EVALITA 2007: THE PART-OF-SPEECH TAGGING TASK*

FABIO TAMBURINI


## SOMMARIO/*ABSTRACT*

Questo contributo descrive il task relativo al *PoS-tagging* in EVALITA 2007, un'iniziativa per la valutazione di sistemi per l'annotazione automatica delle parti del discorso per la lingua italiana. Un numero rilevante di studiosi ha partecipato alla valutazione, sperimentando i vari sistemi sui dati forniti dagli organizzatori. I risultati sono molto interessanti e le prestazioni raggiunte da tali sistemi sono molto alte, specialmente se confrontate con quelle ottenute allo stato dell'arte relativamente alla lingua inglese.

*This paper reports on EVALITA 2007 PoS-tagging task, an initiative for the evaluation of automatic PoS-taggers for Italian. A noticeable number of scholars and teams across Europe participated experimenting their systems on the data provided by the task organisers. The results are very interesting and overall performances are very high, when compared with tagging accuracy for other more studied languages. In particular, the best scores are very close to the state-of-the-art performances obtained for English.*


**Keywords:** NLP, PoS-tagging, Evaluation.

## 1 Introduction

Inside EVALITA 2007, a new initiative for the evaluation of natural language processing tools for Italian language, we organised and managed the task devoted to the evaluation of Part-of-Speech (PoS) taggers. We provided a common framework for the evaluation of tagging systems in a consistent way, supplying the participants with manually annotated data as well as a scoring program for developing and evaluating their systems.

This paper reports on the task organisation and discusses the evaluation results.

### 1.1 Data description

The data sets provided by the organisation were composed of various documents belonging mainly to journalistic and narrative genres, with small sections containing academic and legal/administrative prose. Two separate data sets were provided: the Development Set (DS), composed of 133,756 tokens, was used for system development and for the training phase, while a Test Set (TS), composed of 17,313 tokens, was used as a gold standard for systems evaluation. The ratio between DS and TS is 8/1.

These data have been manually annotated assigning to each token its lexical category (PoS-tag) with respect to two different tagsets producing two different subtasks.

The task organisation did not distribute any lexicon resource with EVALITA data. Each participant was allowed to use any available resource or could freely induce it from the training data.

I wish to thank C. Seidenari for its invaluable help in preparing the EVALITA 2007 data sets.

Table 1: The research teams participating to the task.

| Research Team | Affiliations |
|---|---|
| E. Pianta, R. Zanoli | Foundat. B. Kessler - IRST, Trento, Italy |
| A. Lenci | ILC-CNR and University of Pisa, Italy |
| S. Romagnoli | CILTA, University of Bologna, Italy |
| F. Tamburini | DSLO, University of Bologna, Italy |
| N. Deha, *et al.* | University of Pisa and Synthema Srl, Italy |
| J. Bos, M. Nissim | University of Rome "La Sapienza", University of Bologna, Italy |
| M. Schiehlen | IMS, University of Stuttgart, Germany |
| M. Baroni, *et al.* | Universities of Trento, Stuttgart and Bologna at Forlì, Italy and Germany |
| L. Lesmo | University of Turin, Italy |
| R. Delmonte | University of Venice, Italy |
| M. Ciaramita, J. Atserias | Yahoo! Research, Barcelona, Spain |

#### 1.1.1 Tagsets

PoS-tagging task involved two different tagsets, used to classify the DS data and to be used to annotate TS data. The structure and the principles underlying the tagset design are crucial, both for a coherent approach to lexical classification and to obtain better performance results with

Table 2: The main features of participating systems.

| SYSTEM | Core methods | Lexical resources (other than DS) | U.W. handling methods |
|---|---|---|---|
| FBKirst_Zanoli_POS | Support Vector Machines. | MorphAn., Gazetteers of locations, person and organization names. | None. |
| ILCcnrUniPi_Lenci_POS | Two combined Maximum Entropy taggers. | MorphAn. (100,000 lemmas). | One specific tagger. |
| UniBoCILTA_Romagnoli_POS | HMM. | MorphAn. (35,000 lemmas). | Linear Successive Abstraction and handwritten heuristics. |
| UniBoDSLO_Tamburini_POS | Stacked HMM-TBL. | MorphAn. (120,000 lemmas). | Simple heuristics. |
| UniPiSynthema_Deha_POS | Combination of statistical and rule based methods. | MorphAn. (43,000 lemmas) + specific terminolog. dictionaries. | Specific rules for proper names. |
| UniRoma1_Bos_POS | TnT+CnC. combined with Timbl. | None. | None. |
| UniStuttIMS_Schielen_POS | Support Vector Machines. | Noun/verb/adjective lexicon extracted from Wikipedia. | A special model for unknown words. |
| UniTn_Baroni_POS | TreeTagger. | MorphAn. (35,000 lemmas) + other lexica. | Hand-written rules. |
| UniTo_Lesmo_POS | Hand-written disambiguation rules. | Dictionary (25.000 lemmas) + noun lists and MWE FSA. | None. |
| UniVe_Delmonte_POS | Rule-Based and Statistically Driven. | MorphAn. using a large number of lexica. | Guesser on longer suffixes and/or prefixes+suffixes. |
| Yahoo_Ciaramita_POS_s1 | HMM trained with regularized perceptron algorithm. | None | Prefix/suffix analysis. |
| Yahoo_Ciaramita_POS_s2 | System1 + all feature bigrams. | None | Prefix/suffix analysis. |

automatic techniques, thus they deserve a further discussion.

Italian is one of the languages for which a set of annotation guidelines has been developed in the context of the EAGLES project [1]. Several research groups have been working on PoS annotation to develop Italian treebanks, such as VIT (Venice Italian Treebank [2]) and TUT (Turin University Treebank [3]) and morphological analysers such as the one by XEROX. A comparison of the tagsets used by these groups with EAGLES guidelines reveals that, although there is general agreement on the main parts of speech to be used, considerable divergence exists as regards to the actual classification of Italian words with respect to them. This is the main problematic issue, reflected also in the considerable classification differences operated by the Italian dictionaries.

For the reasons briefly outlined above, we decided to propose two different subtasks for the PoS-tagging evaluation campaign, the first using a traditional tagset (EAGLES-like), the second using a structurally different tagset (DISTRIB). This will allow us to compare different approaches and will give some points to open a discussion on tagset definition, a topic that we believe crucial in the PoS-tagging process.

We refer to the task guidelines [4] for an in-depth discussion of the two tagsets proposed for EVALITA 2007.

#### 1.1.2 Tokenisation issues

The problem of text segmentation (tokenisation) is a central issue in PoS-taggers comparison and evaluation. In principle every system could apply different tokenisation rules leading to different outputs. In this first evaluation campaign we did not have the possibility of handling different tokenisation schemas and following the complex realignment work proposed, for example, inside the GRACE evaluation project [5]. All the development and test data were provided in tokenised format, one token per line followed by its tag.

Participants were requested to return the test set using the same tokenisation format, containing exactly the same number of tokens.

### 1.2 Evaluation metrics

The evaluation was performed in a “black-box” approach: only the systems’ outputs were evaluated. The evaluation metrics were based on a token-by-token comparison and only one tag was allowed for each token. The considered metrics were:

a) *Tagging Accuracy*, defined as the number of correct PoS-tag assignments divided by the total number of tokens in TS.

b) *Unknown Words Tagging Accuracy*, defined as the Tagging Accuracy restricting the computation to unknown words. In this context, for “unknown word” we meant a token present in TS but not in the DS. This, in our opinion, could allow a finer evaluation on the most fruitful morphological techniques or heuristics used to manage unknown words for Italian, a typical challenging problem for automatic taggers.

## 2 Participating systems

Eleven systems completed all the steps in the evaluation procedure and their outputs were officially submitted for this task by their developers.

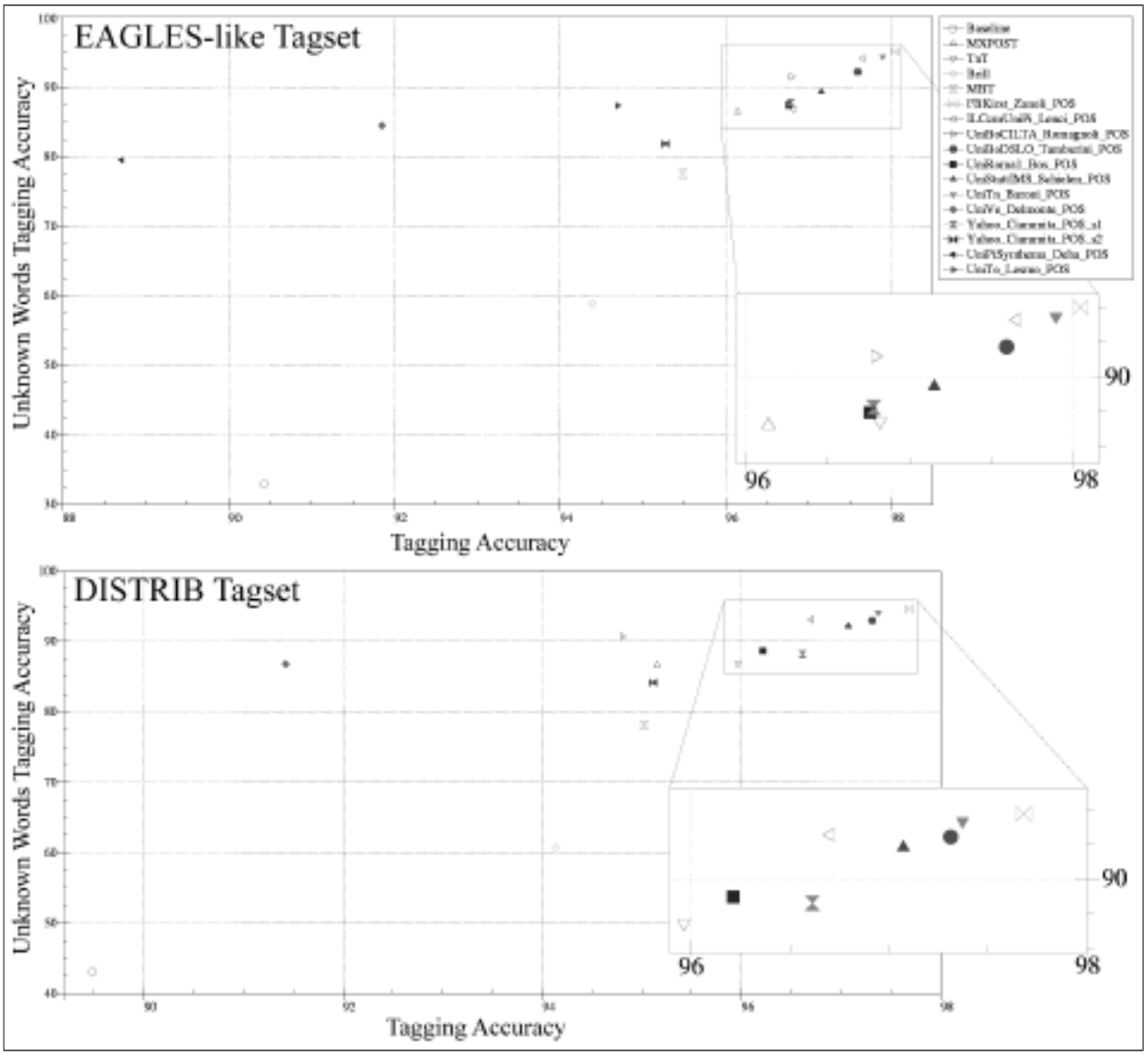


Figure 1: Results obtained by the participating systems.

Table 1 shows the scholars and research teams participating to the evaluation campaign with their affiliations while table 2 describes, in a concise way, the main features of the various systems as communicated by the participants filling a short questionnaire (see detailed descriptions in the participants' papers).

## 3 Results and Discussion

Figure 1 and table 3 show the global results of EVALITA 2007 PoS tagging task for both tagsets, displaying the systems' performances with respect to the proposed metrics.

A baseline algorithm, that assigns the most frequent tag for each known word and the absolute most frequent tag for unknown words, and some well known freely-available PoS-taggers (Brants' TnT [6], Brill's TBL tagger [7], Ratnaparkhi's Maximum Entropy tagger [8], Daelemans' *et. al.* Memory Based tagger [9]) have been inserted into the evaluation campaign as references for comparison purposes. All these taggers were tested by the organisers using the standard configurations described in the respective documentations. No specific optimization options were applied, thus their results represent the basic performances of these systems and could be increased trying to optimise them using their available configuration features.

Examining the systems' performances with respect to their structural features depicted in table 2, we can make some tentative observations:

- there is a group of five systems that performs slightly better that the others exhibiting very high scores (97–

98% of Tagging Accuracy), near to the state-of-the-art performances obtained for English, a language on which there is a long tradition of studies for PoS automatic labelling;

- regarding the core methods implemented by the participants, Support Vector Machines seems to perform quite well: both systems using them are in the top five; the same observation holds for the systems obtained combining or stacking different taggers;
- additional lexical resources seems to play a major role in improving the performances: the systems employing morphological analysers based on big lexica and special techniques for unknown word handling reached the top rankings; these results are clear when analysing the scores considering the Unknown Words Tagging Accuracy metric;
- TnT obtains the best results among the considered reference systems: it embodies a standard, though well optimised, second-order HMM method and employs a sophisticated suffix analysis system that, even in absence of a lexical resource, produces good results;
- the performances obtained by the participating systems remained quite stable when changing the tagset: the best systems tend to exhibit a lowering in performances less than 0.5% when applied to the DISTRIB tagset.

Table 3: *Participants' results with respect to Tagging Accuracy (TA) and Unknown Words Tagging Accuracy (UWTA).*

| SYSTEM | EAGLES-like | | DISTRIB | |
|---|---|---|---|---|
| | TA | UWTA | TA | UWTA |
| FBKirst_Zanoli_POS | 98.04 | 95.02 | 97.68 | 94.65 |
| ILCcnrUniPi_Lenci_POS | 97.65 | 94.12 | 96.70 | 93.14 |
| UniBoCILTA_Romagnoli_POS | 96.79 | 91.48 | 94.80 | 90.72 |
| UniBoDSLO_Tamburini_POS | 97.59 | 92.16 | 97.31 | 92.99 |
| UniRoma1_Bos_POS | 96.76 | 87.41 | 96.21 | 88.69 |
| UniStuttIMS_Schielen_POS | 97.15 | 89.29 | 97.07 | 92.23 |
| UniTn_Baroni_POS | 97.89 | 94.34 | 97.37 | 94.12 |
| UniVe_Delmonte_POS | 91.85 | 84.46 | 91.42 | 86.80 |
| Yahoo_Ciaramita_POS_s1 | 96.78 | 87.78 | 96.61 | 88.24 |
| Yahoo_Ciaramita_POS_s2 | 95.27 | 81.83 | 95.11 | 84.16 |
| UniPiSynthema_Deha_POS | 88.71 | 79.49 | – | – |
| UniTo_Lesmo_POS | 94.69 | 87.33 | – | – |

## REFERENCES


[1] M. Monachini. ELM-IT: EAGLES Specification for Italian morphosintax Lexicon Specification and Classification Guidelines. *EAGLES Document EAG CLWG ELM IT/F*, 1996.

[2] R. Delmonte. Strutture sintattiche dall'analisi computazionale di corpora di italiano. In A. Cardinaletti, A. and F. Frasnedi (Eds.), *Intorno all'italiano contemporaneo. Tra linguistica e didattica.* pp. 187-220, F. Angeli, Milano, 2004.

[3] C. Bosco, V. Lombardo, D. Vassallo, L. Lesmo. Building a treebank for Italian: a data-driven annotation schema. In *Proc. LREC'2000*, 2000.

[4] F. Tamburini, C. Seidenari. EVALITA 2007. The Italian Part-of-Speech Tagging Evaluation - Task Guidelines. http://evalita.itc.it/tasks/pos.html, 2007.

[5] G. Adda, J. Lecomte, J. Mariani, P. Paroubek, and M. Rajman. The GRACE French Part-of-Speech Tagging Evaluation Task. In *Proc. LREC'98*, 1998.

[6] T. Brants. TnT - A Statistical Part-of-Speech Tagger. In *Proc. 6th Conf. on Applied Natural Language Processing*, pp. 224-231, 2000.

[7] E. Brill. Some Advances in Transformation-Based Part of Speech Tagging. In *Proc. 12th National Conference on Artificial Intelligence*, pp. 722-727, 1994.

[8] A. Ratnaparkhi. A Maximum Entropy Model for Part-of-Speech Tagging. In *Proc. EMNLP'96*, pp. 133-142, 1996.

[9] W. Daelemans, J. Zavrel and S. Berck. MBT: A MemoryBased Part of Speech Tagger-Generator. In *Proc. 4th Workshop on Very Large Corpora*, pp. 14-27, 1996.


## 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.