Overview of the EVALITA 2009 Part-of-Speech Tagging Task

Giuseppe Attardi and Maria Simi

Dipartimento di Informatica, Università di Pisa, largo B. Pontecorvo 3, I-56127 Pisa, Italy
{attardi,simi}@di.unipi.it

Abstract. We report on the Evalita 2009 PoS Tagging task, an initiative for the evaluation of automatic PoS Taggers for Italian. The challenge of this year' task consisted in dealing with a large tag set, including morphological traits, and the fact that the training data were extracted from a newspapers corpus while the test data were from a different domain, the Italian Wikipedia. Considering these difficulties, the performance of the participating systems is quite high, compared with state of the art taggers for other languages, with small differences in accuracy between the systems.

Keywords: NLP, PoS Tagging, Evaluation.

1 Motivation

Part of speech tagging might be considered an easy task. Experiments reported in [1] show that state of the art taggers for English, based on machine learning techniques like dependency networks [2], perceptrons [3], SVM [4] or HMM [5], can achieve accuracies above 97%. The reference tag set for English is the Penn Treebank tag set [6], which consists of 36 categories.

The PoS tagging task in Evalita 2007 [7] involved two tagging schemas: an EAGLES compliant tag set, consisting of 32 categories, and a DISTRIB tag set, consisting of 12 categories. Participants could use external resources and lists of multi-words and abbreviations were provided to them. The best submission achieved a remarkable accuracy of around 98%.

For the 2009 edition of the Evalita competition, the task was made more challenging by the combination of two factors:

1. a larger tag set, consisting 37 tags with morphological variants, resulting in 336 different morphed tags;
2. domain adaptation: the training corpus consists of newspaper articles from La Repubblica, while the development and test set are extracted from the Italian Wikipedia.

The evaluation aimed at verifying how much these additional complexities would influence accuracy. Separate scores are hence provided for accuracy with respect to both the morphed tags and the 37 tags without morphology; a separate evaluation is also provided for unknown words, since the domain shift makes this aspect especially critical.

2 Task definition

The evaluation was based on three data sets provided by the organizers: a Training Set containing data annotated using the Tanl tagset [12]; a Dev Set to be used for testing the systems during development; a Test Set, containing blind test data for the evaluation.

The corpora were annotated using the version of the Tanl tagset that includes morphological features and consists of 336 tags, grouped into 14 basic categories. The task measures the ability of taggers to handle a large tagset, and hence the possibility of using a POS tagger to obtain both lexical and morphological information, possibly without resorting to external lexicons or other resources.

The evaluation is organized in two subtasks:

1. a *closed task*, where participants are not allowed to use any external resources besides the supplied Training and Dev Sets.
2. an *open task*, where participants can use external resources.

The participants were asked to submit up to four runs of their systems, since we wanted to encourage comparison among different approaches and experiments.

3 Corpora description

The training corpus consists of articles from the online edition of the newspaper La Repubblica (http://www.repubblica.it/). The corpus consists in 108,874 word forms divided into 3,719 sentences.

These data have been annotated in several steps: the first step was performed by the group of Andrea Baroni at the Università di Bologna and consisted in manually assigning a set of coarse-grain POS tags; then the MorphIt! [9] tool was used to suggest a list of possible morphological tags for each token; then the correct one was handpicked; then a conversion script incorporating some heuristics was used to convert the POS and morphological tags into the Tanl tagset.

The corpus was manually revised and then automatically cross-checked with an Italian lexicon of over 1.25 million forms, in order to identify anomalies.

Both the Dev and the Test sets were extracted from the Italian Wikipedia. This collection was chosen to provide a good test for domain adaptation, since Wikipedia articles are quite different in style and terminology from the newspaper domain of the training corpus. Wikipedia also uses different conventions for punctuation; for example double quotes are quite frequent in the Wikipedia but were totally absent from the training corpus.

The Dev Set consists of 5021 word forms divided into 147 sentences. The number of new words in the Dev Set with respect to the Training Set is quite high (870/5021, i.e. over 17%), since Wikipedia articles cover many different topics, involving the use of specialized terminologies.

The Test Set consists of 5066 word forms divided into 147 sentences. The percentage of words not present in the Training Set is still around 17%.

3.1 Data format

The training corpus is provided as a single file, UTF-8 encoded, in tokenized format, one token per line followed by its tag, separated by a TAB. Here is an example:

```
A             E
ben           B
pensarci      Vfc
,             FF
l'            RDns
intervista    Sfs
dell'         EAns
on.           SA
Formica       SP
è             VAip3s
stata         VApsfs
accolta       Vpsfs
in            E
genere        Sms
con           E
disinteresse  Sms
.             FS
```

The example illustrates some tokenization issues:

1. abbreviations are properly identified as tokens (`on.`);
2. apostrophes representing a truncation are kept with the truncated token (`l'intervista`);
3. possible multi-word expressions (`in_genere`) are not combined into a single token;
4. clitics are not separated from the token (`pensarci`).

The Tanl tagset was designed according to the EAGLES guidelines [10], an agreed standard in the NLP community. In particular it was derived from the morphosyntactic classification of the ISST corpus [11].

Tanl provides three levels of POS tags: coarse-grain, fine-grain and morphed tags. The coarse-grain tags consist of the 14 categories listed in Table 1.

Table 1. Coarse-grain tags.

Tag	*Description*	*Tag*	*Description*
A	adjective	N	numeral
B	adverb	P	pronoun
C	conjunction	R	article
D	determiner	S	noun
E	preposition	T	predeterminer
F	punctuation	V	verb
I	interjection	X	residual class

Table 2 presents the list of fine-grain tags (37), with short descriptions.

Table 2. Fine-grain tags.

Tag	*Description*	*Tag*	*Description*
A	adjective	NO	ordinal number
AP	possessive adjective	PC	clitic pronoun
B	Adverb	PD	demonstrative pronoun
BN	negation adverb	PE	personal pronoun
CC	coordinative conjunction	PI	indefinite pronoun
CS	subordinative conjunction	PP	possessive pronoun
DD	demonstrative determiner	PQ	interrogative pronoun
DE	exclamative determiner	PR	relative pronoun
DI	indefinite determiner	RD	determinative article
DQ	interrogative determiner	RI	indeterminative article
DR	relative determiner	S	common noun
E	preposition	SA	abbreviation
EA	articulated preposition	SP	proper noun
FB	balanced punctuation	T	predeterminer
FC	clause boundary punct.	V	main verb
FF	comma, hyphen	VA	auxiliary verb
FS	sentence boundary punct.	VM	modal verb
I	interjection	X	residual class
N	cardinal number		

The morphed tags consist of 336 categories, which include morphological information encoded as follows:

gender: *m* (male), *f* (female), *n* (underspecified)
number: *s* (singular), *p* (plural), *n* (underspecified)
person: 1 (first), 2 (second), 3 (third)
mode: *i* (indicative), *m* (imperative), *c* (conjunctive), *d* (conditional), *g* (gerund), *f* (infinite), *p* (participle)
tense: *p* (present), *i* (imperfect), *s* (past), *f* (future)
clitic: *c* marks the presence of agglutinative clitics.

The set of morphed Tanl tags used for the EVALITA09 POS tagging subtask is described in detail at http://medialab.di.unipi.it/wiki/index.php/Tanl_POS_Tagset.

Of the 336 possible morphed tags, only 234 were indeed present in the training corpus, since some of the legal combinations are quite rare.

4 Evaluation measures

The evaluation is performed with a "black box" approach: only the system output is evaluated. Only one tag is allowed for each token, and the evaluation metrics are based on a token-by-token comparison. The following metrics are computed:

1. *Tagging accuracy* (TA): it is defined as the percentage of correctly tagged tokens with respect to the total number of tokens in the Test Set.

2. *Unknown Words Tagging Accuracy* (UWTA): it is defined as the tagging accuracy restricted to the unknown words. In this context "unknown word" means a token present in the Test Set but not in the Training Set.

To measure the loss in accuracy due to morphology, TA was measured both with respect to the whole tag set, including all the morphological variants of the fine-grain tags of Table 2 (POS), and also with respect to the fine-grained tags without morphology (CPOS). Evaluation was performed by a script made available to the participants, which computes both the overall accuracy and the error rate for each tag.

5 Participation results

The 8 teams listed in **Table 3**, out of the 15 who had expressed interest, participated in the Evalita 2009 PoS tagging task by actually submitting runs of their systems.

Table 3. Teams participating in the Evalita 2009 PoS Tagging task.

Research Team	*Main investigator*	*Affiliation*
SemaWiki	G. Attardi	Dip. di Informatica, Univ. di Pisa, Italy
CST_Søgaard	A. Søgaard	Centre for Lang. Tech., Univ. of Copenhagen, Denmark
Gesmundo	A. Gesmundo	Università di Genova, Italy
Felice-ILC	F. Dell'Orletta	ILC-CNR, Pisa, Italy
Lesmo	L. Lesmo	Dip. di Informatica, Univ. di Turin, Italy
Pianta	E. Pianta	Found. B. Kessler – IRST, Trento, Italy
Rigutini	L. Rigutini	Dip. di Ing. Informatica, Univ. di Siena, Italy
Tamburini	F. Tamburini	DSLO, Università di Bologna, Italy

Table 4. Open task results.

Team	*POS TA*	*CPOS TA*	*POS UWTA*	*CPOS UWTA*	*Rank*
SemaWiki 2	**96.75%**	**97.03%**	**94.62%**	**95.30%**	**1**
SemaWiki 1	96.44%	96.73%	94.27%	95.07%	2
SemaWiki 4	96.38%	96.67%	93.13%	93.81%	3
SemaWiki 3	96.14%	96.42%	92.55%	93.24%	4
Pianta	96.06%	96.36%	92.21%	93.24%	5
Lesmo	95.95%	96.26%	92.33%	93.01%	6
Tamburini 1	95.93%	96.40%	90.95%	92.67%	7
Tamburini 2	95.63%	96.16%	91.07%	92.78%	8

By comparing POS and CPOS accuracy, we observe that the drop in accuracy due to errors in morphology is lower than 0.30% in general, and close to 1% in the case of unknown words. If we restrict the attention to the tagging of unknown words, the loss in accuracy with respect to the general case ranges from 1.4% to 3%.

Table 5 reports the results obtained by the participating teams in the Closed Task.

Table 5. Closed task results.

Team	*POS TA*	*CPOS TA*	*POS UWTA*	*CPOS UWTA*	*Rank*
Felice_ILC	**96,34%**	**96,91%**	91,07%	93,36%	**1**
Gesmundo	95,85%	96,48%	**91,41%**	**93,81%**	2
SemaWiki 2	95,73%	96,52%	90,15%	93,47%	3
SemaWiki 1	95,24%	96,00%	87,40%	90,72%	4
Pianta	93,54%	94,10%	85,45%	87,74%	5
Rigutini 2	93,37%	94,15%	86,03%	88,43%	6
Rigutini 3	93,31%	94,15%	86,03%	88,55%	7
Rigutini 4	93,29%	94,17%	85,34%	88.09%	8
Rigutini 1	93,10%	93,76%	84,54%	87,06%	9
CSTSøgaard 1	91,90%	93,21%	86,03%	89,58%	10
CSTSøgaard 2	91,64%	93,21%	86,14%	89,92%	11

When no external resources are used, the loss in performance due to the prediction of the morphology is lower than 1% in general, but close to 3% in the case of unknown words. More difficult appears, for all the systems, the tagging of unknown words: the loss of accuracy ranges from 4.4% to as much as 8.5%.

6 Summary of the Approaches

In order to provide a quick comparison of the techniques used in the submissions, we asked each team to report some summary information about their systems.

Except for Lesmo and Gesmundo, all participants used a combination of taggers. The component taggers were usually statistical POS taggers, except for two cases that used TBL in a combination. A few teams developed their own tagger (Lesmo, Gesmundo, Felice-ILC) while most used readily available tools.

Except for Tamburini and SemaWiki who used special heuristics, no special handling for unknown words was reported. Further details are collected in **Table 6**.

Table 6. Summary of approaches.

Team	*Type*	*Components*	*Model order*	*n-gram*	*Train. Feat.*	*Search*
SemaWiki	Combination or cascade + rules	Hunpos, TreeTagger	Second	3-gram		Viterbi
CST_Søgaard	Combination	Brill, TreeTagger, MaxEntropy + combination classifier		1-gram in classifier		none
Gesmundo	Single	Perceptron	First	2-gram bidirectional	515k	beam-search
Felice-ILC	Combination	HMM, SVM, MaxEntropy	Second	4-gram	SVM: 91400, ME: 939000	Viterbi
Lesmo	Rule based	573 rules				
Pianta Rigutini	Cascade of 4 classifiers	SVM	First	varying		Viterbi
Tamburini	Combination	HMM, TBL		2-gram, 3-gram		Viterbi; Brill algorithm

In the Open Task the following resources were used:

Table 7. External resources.

Team	*Lexicon size*	*Gazetteer size*
SemaWiki	65,000 lemmas / 1,200,000 forms	0
Pianta		74,000
Lesmo	*n.a.*	
Tamburini	120,000 lemmas	

7 Error Analysis

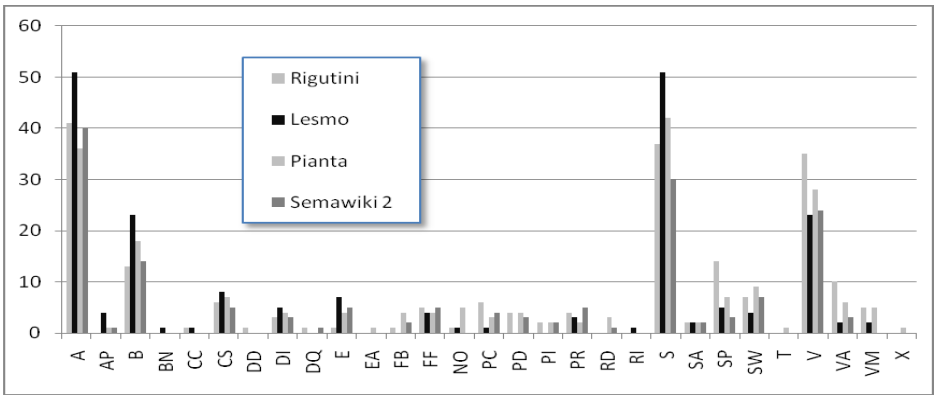

Figure 1. Distribution of errors.

Figure 1 Figure 1. Distribution of errors.shows the number of errors, grouped by fine-grained POS categories, of the best run of each team in the Open Task. The classes were more errors occur are those of adjectives, nouns and verbs, which are also the most frequent, with respectively 3528, 12440 and 6650 occurrences in the test set. The analysis for the closed task produces a similar distribution.

8 Conclusions

Considering the sources of complexity introduced by the Evalita 2009 PoS Task, i.e. large tag set and domain adaptation, the accuracy of the participating systems is quite high, as compared to state of the art taggers for other languages. In the Open Task there is only a 1.12% difference between the worst and the best performing system. In the Closed Task the difference in performance is higher but the three top performing systems are quite close. These results may be interpreted as showing that aiming for a rich grammatical tagging is feasible and practical for Italian and that these tools can be reused across domains.

Acknowledgments. We are grateful to all the SemaWiki team for help in producing high quality resources for this task and in particular to Simonetta Montemagni for supervising the linguistic soundness and providing advice on critical cases. The resources for the Evalita 2009 POS task were developed within project SemaWiki, partially funded by the Fondazione Cassa di Risparmio di Pisa.

References

1. Tsuruoka, Y., Tsujii, J.: Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data. In: Proceedings of HLT-EMNLP, pp. 467--474 (2005)

2. Toutanova, K., Klein, D., Manning, C., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of HLT-NAACL, pp. 505--512 (2003)
3. Collins, M.: Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In: Proceedings of EMNLP, pp. 1--8 (2002)
4. Gimenez, J., Marquez, L.: Fast and accurate part-of-speech tagging: the SVM approach revisited. In: Proceedings of RANLP, pp. 158--165 (2003)
5. Brant, T.: TnT–a statistical part-of-speech tagger. In: Proceedings of the 6th Applied NLP Conference (2000)
6. Marcus, M. P., Santorini, B., Marcinkiewicz, M. A.: Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics, vol. 19, issue 2, pp. 313--330 (1993).
7. Tamburini, F., Evalita 2007: The Part-of-speech Tagging Task. IA-Intelligenza Artificiale, Anno IV, issue 2, pp. 4--7 (2007)
8. De Mauro, T.: Il dizionario della lingua italiana, http://www.demauroparavia.it/ (2007)
9. Zanchetta, E., Baroni, M.: Morph-it! A free corpus-based morphological resource for the Italian language. In: Proceedings of Corpus Linguistics, http://dev.sslmit.unibo.it/linguistics/morph-it.php (2005)
10. Monachini, M.: ELM-IT: An Italian Incarnation of the EAGLES-TS. Definition of Lexicon Specification and Classification Guidelines. Technical report, Pisa (1995)
11. Montemagni, S. et al.: Building the Italian Syntactic-Semantic Treebank. In: Abeillé (ed.), Building and using Parsed Corpora, Language and Speech series. Kluwer, Dordrecht, pp. 189--210 (2003)
12. Attardi, G., et al.: Tanl - Text Analytics and Natural Language processing: Analisi di Testi per il Semantic Web e il Question Answering, http://medialab.di.unipi.it/wiki/SemaWiki/