EVALITA 2009

Italian Part-Of-Speech Tagging Evaluation

Task Guidelines

Giuseppe Attardi and Maria Simi
Dipartimento di Informatica, Università di Pisa, Italy
attardi@di.unipi.it

1. Introduction

The following are the guidelines for the PoS-tagging task of the EVALITA 2009 evaluation campaign.

The evaluation will be based on two data sets provided by the organizers: the first, referred to as **Training Corpus (TrC)** contains data annotated using the *Tanl* tagset [5] and must be used for training participating systems; the second, referred to as **Test Set (TeS)** contains blind test data for the evaluation.

The corpora are annotated using the version of the Tanl tagset that includes morphological features and consists of 328 tags, from 14 basic categories. The task will hence evaluate the ability of taggers to handle a large tagset, useful for obtaining both lexical and morphological information from a POS tagger.

There will be two subtasks:

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

Participants are required to provide a brief description of their system and a full notebook paper describing their experiments, in particular the techniques, the resources used and presenting an analysis of the results.

2. Corpora Description

The data sets provided by the organizers consist of articles from the online edition of the newspaper La Repubblica (http://www.repubblica.it/).

The whole 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! [2] automated tool was used to assign a list of possible morphological tags to each token; a conversion script incorporating some heuristics was used to convert the POS and morphological tags into the Tanl tagset.

A final manual revision was applied to the whole corpus followed by a complete automated cross-check with an Italian lexicon of over 1,25 million forms.

Participants are not allowed to distribute the EVALITA 2009 data as stated in the non-disclosure agreement (licence) signed before receiving the data.

3. Data Format

The training corpus will be provided as a single Unix file, UTF-8 encoded, in tokenized format, one token per line followed by its tag, separated by a TAB, according to the following schema:

```
<TOKEN_1> <TAG1>
<TOKEN_2> <TAG2>
...
<TOKEN_N> <TAGN>
<EMPTY LINE>
```

An empty line terminates each sentence. 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:

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

The blind version of TeS will contain just non annotated tokens, one per line.

The format of the submitted run files must be the same as TrC, containing one token per line with the corresponding predicted POS tag. The evaluation script will compare the submitted run files with the reference file line-by-line, hence a misaligned file will fail the evaluation.

4. The Tanl Tagset

The Tanl tagset is designed according to the EAGLES guidelines [3], an agreed standard in the NLP community. In particular it was derived from the morpho-syntactic classification of the ISST corpus [4].

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

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

The fine-grain tags (36) are reported, defined and exemplified in the following table:

Fine-grain tag	Description	Examples	Contexts of use
A	adjective	*bello, buono, pauroso, ottimo*	una **bella** passeggiata un **ottimo** attaccante una persona **paurosa**
AP	possessive adjective	*mio, tuo, nostro, loro*	a **mio** parere il **tuo** libro
B	adverb	*bene, fortemente, malissimo, domani*	arrivo **domani** sto **bene**
BN	negation adverb	*non*	**non** sto bene
CC	coordinative conjunction	*e, o, ma*	i libri **e** i quaderni vengo **ma** non rimango
CS	subordinative conjunction	*mentre, quando*	**quando** ho finito vengo **mentre** scrivevo ho finito l'inchiostro
DD	demonstrative determiner	*questo, codesto, quello*	**questo** denaro **quella** famiglia
DE	exclamative determiner	*che, quale, quanto*	**che** disastro! **quale** catastrofe!
DI	indefinite determiner	*alcuno, certo, tale, parecchio, qualsiasi*	**alcune** telefonate **parecchi** giornali **qualsiasi** persona
DQ	interrogative determiner	*cui, quale*	i **cui** libri
DR	relative determiner	*che, quale, quanto*	**che** cosa **quanta** strada **quale** formazione
E	preposition	*di, a, da, in, su, attraverso, verso, prima_di*	**a** casa **prima_di** giorno **verso** sera

Fine-grain tag	Description	Examples	Contexts of use
EA	articulated preposition	*alla, del, nei*	**nel** posto
FB	balanced punctuation	*( ) [ ] { } - _ ` ' ‘ ’ " “ ” « »*	**(**sempre**)**
FC	clause boundary punctuation	*. - : ;*	segue**:**
FF	comma, hyphen	*, -*	carta**,** penna 30**-**40 persone
FS	sentence boundary punctuation	*. ? ! ...*	cosa**?**
I	interjection	*ahimè, beh, ecco, grazie*	**Beh**, che vuoi?
N	cardinal number	*uno, due, cento, mille, 28, 2000*	**due** partite **28** anni
NO	ordinal number	*primo, secondo, centesimo*	**secondo** posto
PC	clitic pronoun	*mi, ti, ci, si, te, ne, lo, la, gli*	**me ne** vado **si** sono rotti **mi** lavo **gli** parlo
PD	demonstrative pronoun	*questo, quello, costui, ciò*	**quello** di Roma **costui** uccide
PE	personal pronoun	*io, tu, egli, noi, voi*	**io** parto **noi** scriviamo
PI	indefinite pronoun	*chiunque, ognuno, molto*	**chiunque** venga i diritti di **ognuno**
PP	possessive pronoun	*mio, tuo, suo, loro, proprio*	il **mio** è qui più bella della **loro**
PQ	interrogative pronoun	*che, chi, quanto*	non so **chi** parta **quanto** costa? **che** ha fatto ieri?
PR	relative pronoun	*che, cui, quale*	**ciò** che dice il **quale** afferma a **cui** parlo
RD	determinative article	*il, lo, la, i, gli, le*	**il** libro **i** gatti
RI	indeterminative article	*uno, un, una*	**un** amico **una** bambina
S	common noun	*amico, insegnante, verità*	l’**amico** la **verità**
SA	abbreviation	*ndr, a.C., d.o.c., km*	30 **km** sesto secolo **a.C.**
SP	proper noun	*Monica, Pisa, Fiat, Sardegna*	**Monica** scrive
T	predeterminer	*tutto, entrambi*	**tutto** il giorno **entrambi** i bambini
V	main verb	*mangio, passato, camminando*	**mangio** la sera il peggio è **passato** ho **scritto** una lettera

Fine-grain tag	Description	Examples	Contexts of use
VA	auxiliary verb	*avere, essere, venire*	il peggio **è** passato **ho** scritto una lettera **viene** fatto domani
VM	modal verb	*volere, potere, dovere, solere*	non **posso** venire **vuole** comprare il libro
X	residual class	it includes formulae, unclassified words, alphabetic symbols and the like	distanziare di **43"** mi **piacce**

The morphed tags consist of 328 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>.

5. Evaluation Metrics

The evaluation is performed in a "black box" approach: only the system output is evaluated. The evaluation metrics will be based on a token-by-token comparison and only ONE tag is allowed for each token.

The considered metrics will be:

a) *Tagging accuracy*: it is defined as the percentage of correctly tagged tokens with respect to the total number of tokens in TeS.

b) *Unknown Words Tagging Accuracy*: it is defined as the *Tagging Accuracy* restricting the computation to unknown words. In this context "unknown word" means a token present in TeS but not in TrC.

The results of a baseline PoS-tagger (TnT) will be used as reference for comparison purposes.

Evaluation will be performed by the script `evalitaPos.py` that will be provided to participants.

6. Evaluation Details

For important dates and deadlines we refer to the official Evalita 2009 site http://evalita.fbk.eu/important_dates.html .

Participants should submit the results of their runs sending to the organizers email address (`evalita@di.unipi.it`)a file in the same format as the Training Corpus, named as:

- `EVALITA09_POS_Closed_ParticipantName` (for submission to the closed subtask)
- `EVALITA09_POS_Open_ParticipantName` (for submission to the open subtask)

Only one result file for each subtask will be accepted.

After the submission deadline the organizers will evaluate the submitted runs and will send to each participant the score of his submissions as well as the *gold-standard* version of TeS.

7. References

[1] T. De Mauro. 2007. *Il dizionario della lingua italiana*. On-line version http://www.demauroparavia.it/

[2] E. Zanchetta, M. Baroni. 2005. Morph-it! A free corpus-based morphological resource for the Italian language. *Proc. of Corpus Linguistics 2005*, University of Birmingham, Birmingham, UK. http://dev.sslmit.unibo.it/linguistics/morph-it.php

[3] M. Monachini. 1995. ELM-IT: An Italian Incarnation of the EAGLES-TS. Definition of Lexicon Specification and Classification Guidelines. Technical report, Pisa.

[4] S. Montemagni, et al. 2003. Building the Italian Syntactic-Semantic Treebank. In Abeillé (ed.), *Building and using Parsed Corpora, Language and Speech series*, Kluwer, Dordrecht, 189–210.

[5] G. Attardi et al. 2008. 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.