EVALITA 2007
THE ITALIAN PART-OF-SPEECH TAGGING EVALUATION
- TASK GUIDELINES -

Fabio Tamburini, Corrado Seidenari
Dipartimento di Studi Linguistici e Orientali, Università di Bologna, Italy.
fabio.tamburini@unibo.it, corrado.seidenari2@unibo.it

1. INTRODUCTION

The following are the guidelines for the PoS-tagging task of EVALITA 2007 evaluation campaign.

Participants to the evaluation task are required to use the data provided by the organization to set up their systems. The organisation will provide two data sets: the first, referred to as **Development Set (DS)** contains data manually classified using two different tagsets (see a following section for tagsets descripton) and must be used to train participants' systems; the second, referred to as **Test Set (TS)** contains the test data for the evaluation.

Participants are allowed to use other resources in their systems, both for training and to enhance final performances, but the results must conform to the proposed tagsets.

Participants are also required to send a brief description of the system, especially considering techniques and resources used, and (if available) a complete bibliographic reference and the full paper in electronic format.

2. DATA DESCRIPTION

The data sets provided by the organisation are composed of various documents belonging mainly to journalistic and narrative genres, with small sections containing academic and legal/administrative prose.

These data have been manually annotated assigning to each token its lexical category (PoS-tag) with respect to two different tagsets generating two different subtasks. Each participant is required to use both tagsets and, consequently, annotate the TS using them and provide the organization with two resulting output file that will be used for evaluation. The two subtask will be evaluated separately. Even if it requires more work we encourage the participants to apply their systems to both subtasks because they involve quite different tagsets designed with different criteria, as you will see in the following sections. The comparison of the final results is expected to rise interesting issues. Please, let us know if you find some errors in the data annotations; this will allow us to update and redistribute them to the participants in a enhanced form.

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

We do not distribute a lexicon resource with EVALITA07 data. Each participant is allowed to use any available resource or can freely induce it from the training data.

2.1. Tagsets

PoS-tagging task involves two different tagsets, used to classify the DS data and to be used to annotate TS data. We believe that 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 automatic techniques, and deserve a further discussion.

The Italian reference grammars examined are: Serianni [1989], which can be considered an authoritative work among the traditional grammars of Italian, and Renzi et al. [1988, 1991, 1995], an innovative work within the framework of generative grammar. We also refer to Sensini [1997] and Dardano/Trifone [1997].

The reference dictionaries are: De Mauro [1999], the most comprehensive lexicographic work on Italian in use, and Sabatini/Coletti [1997], which takes an innovative approach to the problem of categorisation. We also referred to Zingarelli [2000] and Devoto/Oli [2001].

Serianni, who adopts a traditional terminology in his Grammar, proposes a distinction into ten parts of speech (noun, article, adjective, numeral, pronoun, preposition, conjunction, interjection, verb, adverb). However, he admits the problematic nature of the class of conjunctions, prepositions and adverbs.

Renzi *et al*., who assume the principle of the centrality of syntax in their description, starting from the phrase to go down to the parts of speech, adopt some traditional designations such as: name (the head of a noun phrase), article (determiner of a noun or a noun modifier), adjective (head of the adjectival phrase), verb (head of the verbal phrase), adverb (head of the adverbial phrase), preposition (head of a prepositional phrase), pronoun.

Sensini adopts a traditional classification with nine parts of speech, divided into variable (article, name, adjective, pronoun, verb) and invariable (adverb, preposition, conjunction, interjection).

Dardano and Trifone propose the same classification, underlying the vague boundaries of some classes, for example the mixing of the classes of conjunction, preposition and adverb.

With regard to dictionaries the whole picture does not change radically.

For the categorisation of lemmas, the GRADIT dictionary uses nine parts of speech (article, noun, adjective, pronoun, verb, adverb, preposition, conjunction, interjection).

Roughly the same division into nine parts of speech can be found in other dictionaries, with some differences in listing, particularly for the classification of conjunctions and adverbs.

Apart from the traditional categories, the DISC dictionary considers 'textual conjunctions' with their respective locutions. These are elements that cannot be easily included in the nine parts of discourse and are vaguely assigned to the category of conjunctions or adverbs. For some other elements that have a primary function in a sentence, the dictionary indicates any possible use as textual conjunction.

Italian is one of the languages for which a set of annotation guidelines has been developed in the context of the EAGLES project [Monachini, 1995]. Several research groups have worked on PoS annotation to develop treebanks, such as VIT (Venice Italian Treebank [Delmonte, 2004]) and TUT (Turin University Treebank) [Bosco et al., 2000; Bosco, 2003] and morphological analysers such as of XEROX. A comparison of the tag sets used by these groups with Monachini's guidelines reveals that though there is general agreement on the main parts of speech to be used, considerable divergence exists when it comes to the actual classification of Italian words with respect to these main PoS classes. That is the main problematic issue.

The main categories identified within the EAGLES project are nouns, verbs, adjectives, adverbs, determiners, pronouns, articles, adposition, conjunctions, numerals, interjections and residuals. The actual tagset can then be obtained by further subdividing these categories by means of semantics or morpho-syntactic criteria. The tagsets proposed by the cited works differ, however, on the criteria used for subdividing the main classes and hence they are rather different. The classes for which differences of opinion are most evident are adjectives, determiners, conjunctions and adverbs. These differences will then influence the kind of conclusions one can draw from the annotated corpus since they do not boil down to simply terminological differences resolvable by a mere one-to-one relabelling or by mapping different classes into a greater one.

These, and other, factors drove us to propose to semi-automatically induce the word classes. This tagset induction process is described in detail in [Bernardi *et al.* 2005; 2006].

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, the second using a structurally different tagset. This will allow us to compare different approaches and will give some points to open a discussion on tagset definition, a point that we believe crucial in the PoS-tagging process.

2.1.1 Traditional "EAGLES-like" Tagset

The first tagset proposed is designed according to the EAGLES guidelines [Monachini, 1995], one of the few agreed standard in the NLP community. In particular it is similar to the Level 1 of the morpho-syntactic classification proposed by Monachini.

As to the classification mismatches and the actual disagreement in assigning words to PoS classes, we relied on suggestions and instances mainly taken from the online version of the dictionary edited by De Mauro [2007].

This is the complete tag list as used for this EVALITA07 subtask:

Verbs	*V_AVERE*	All forms of verb *avere*.
	V_ESSERE	All forms of verb *essere*.
	V_MOD	All forms of verbs *potere, dovere, volere*.
	V_PP	Past and present participles.
	V_GVRB	General verb forms.
	V_CLIT	Cliticized verb forms (e.g. *andarci*, *dimmelo*).
Nouns	*NN*	Common nouns.
Proper Nouns	*NN_P*	-
Articles	*ART*	-
Prepositions	*PREP*	Simple prepositions.
	PREP_A	Prepositions fused with articles. (*prep. articolate*)
Adjectives	*ADJ*	Qualifying adjectives.
	ADJ_DIM	Demonstrative adjectives.
	ADJ_IND	Indefinite adjectives.
	ADJ_IES	Interrogative or exclamative adjectives.
	ADJ_POS	Possessive adjectives.
	ADJ_NUM	Numeral adjectives.
Conjunctions	*CONJ_C*	Coordinating conjunctions.
	CONJ_S	Subordinating conjunctions.
Adverbs	*ADV*	-
Interjections	*INT*	-
Numbers	*C_NUM*	Cardinal numbers.
Pronouns	*PRON_PER*	Personal pronouns.
	PRON_REL	Relative pronouns.
	PRON_DIM	Demonstrative pronouns.
	PRON_IND	Indefinite pronouns.
	PRON_IES	Interrogative or exclamative pronouns.
	PRON_POS	Possessive pronouns.
Symbols	*NULL*	Codes, delimiters…
Punctuation Marks	*P_EOS*	Full stop ".", exclamative and interrogative marks "!", "?" closing a sentence.
	P_APO	Apostrophe when used as quotation mark.
	P_OTH	Other punctuation marks.

Proper Noun Management

The annotation of named entities (NE) posed a number of relevant problems.

The most coherent way to handle such kind of phenomena is to consider the NE as a unique token assigning to it the NN_P tag. Unfortunately this is not a viable solution for this evaluation task, and, moreover, a lot of useful generalisation on trigram sequences (e.g. *Ministero/dell'/Interno – NN_P/PREP_A/NN_P*) would be lost if adopting such kind of solution.

Anyway, the annotation of sequences like "*Banca Popolare*" and "*Presidente della Repubblica Italiana*" deserve some attention and a clear policy.

For EVALITA07 we decided to annotate as *NN_P*s those words, belonging to the NE, marked with the uppercase letter. Thus the example above, and some others, have been annotated as:

Banca	*NN_P*	*Presidente*	*NN_P*	*Ordine*	*NN_P*	*Accademia*	*NN_P*
Popolare	*NN_P*	*della*	*PREP_A*	*dei*	*PREP_A*	*militare*	*ADJ*
		Repubblica	*NN_P*	*medici*	*NN*	*di*	*PREP*
		Italiana	*NN_P*			*Amburgo*	*NN_P*

Beware that in other cases the uppercase initial has not been considered sufficient to determine a *NN_P*:

...certo numero di casi vengono segnalati anche nei Paesi dove la malaria…

...non si presentava necessariamente in contraddizione con lo Stato sociale…

Foreign words

Non-Italian words are annotated, when possible, following the same criteria adopted for Italian ones.

2.1.2 A Tagset Distributionally/Syntactically Oriented

The tagset lists 14 PoS classes as they were actually induced by the algorithm proposed in [Bernardi *et al.* 2005; 2006]. Each class will be briefly described below.

Nominal	*N*
Verbal	*V*
Adjectival	*ADJ*
Adverbial	*ADV*
Entity	*ENTITIES*
Relative	*REL*
Subordinator	*SUB_ARG*
	SUB_ADJ
Coordinator	*COORD*
Argument-Operator	*ARG_DET*
	ARG_PREP
Prepositional	*PREP_POLI*
	PREP_NA
	PREP_VA

The list above does not include the tags added for residuals, which are:

Punctuation Mark	*PUNT*	Any punctuation mark.
Symbol	*NULL*	Codes, delimiters…

Basically, in the tagset presented, each PoS class results from the generalization of a lexico-syntactic prototype (i. e. a set of sample words and their related syntactic patterns) that was semi-automatically extracted from a dependency treebank.

As a result, the 14 PoS prototypes presented are comparable with the EAGLES main classes only partially. Four prototypes roughly correspond to the ones proposed by EAGLES guidelines: Nominal (≈Noun, Proper Noun), Verbal (≈Verb), Adverbial (≈Adverb), Adjectival (≈Adjective). Significant differences, as will be discussed below, arise with respect to Entity, Relative, Subordinator, Coordinator, Argument-Operator and Prepositional.

- Entity:

Entity (*ENTITIES*) prototype includes **non-functional items engaged in Head-Argument relation with a verb**. Typical Entity class members are pronominals, such as "*coloro*" (those) in the following example:

... tutti coloro che offrono aiuto sono i benvenuti ...

- Relative:

Relative (*REL*) prototype contains mainly **pronominals** and **adverbials** when **engaged in relative adjunctions**.

... ai terreni su cui esistevano diritti ...
... vicino all'università dove nel '90 scoppiò la rivolta ...

- Coordinator:

Coordinator (*COORD*) includes items behaving as **Head**, **bridging two or more structures connected in a non-hierarchical fashion**. Examples are straightforward coordinators such as "*e*" (and), "*o*" (or), "*ma*" (but), etc.

- Subordinator:

Subordinator (*SUB*) prototype includes expressions syntactically behaving as **Head**, **bridging two clauses connected in a hierachical fashion**. In fact, the induction process detected two different PoS prototypes:

a) *SUB_ADJ*, subordinators Head of an clausal Adjunct e.g. "*quando*" (when), "*perché*" (why);
b) *SUB_ARG*, subordinators Head of a clausal Argument, typically dependent on a verbal Head (e.g "*che*" (that), "*di*" (to);

as illustrated by the following examples:

a) ... si applicano anche quando si tratta di togliere un ingombro ...
b) ... salvo che esigenze tecniche impongano di costruirlo ...

- Argument-Operator:

Argumentizer (*ARG*) protoype includes all those expressions distributionally close to determiners, tipically engaged as Head **in argument structures** mainly **dependent on a verbal Head**. This prototype was splitted into two classes:

a) *ARG_DET* roughly including determiners
... il comportamento dei pm ...
... l'unica volta che mio padre mi portò al cinema ...
b) *ARG_PREP* containing prepositions.
... spetta a Massimo D'Alema dire se ...

- Prepositional:

Prepositional (*PREP*) prototype contains **prepositions**, for instance "*attraverso*", "*secondo*", "*con*", "*sul*", "*nel*", "*di*", "*degli*", etc. which, directly or indirectly governing noun structures, **yield verb or noun adjuncts**. There are three different prepositional prototypes:

a) *PREP_POLI*, for prepositions (e.g. "*attraverso*", "*secondo*", "*contro*", etc.) governing determiner or prepositional structures and forming verb adjuncts (this class typically includes the so called polysyllabic prepositions);

... protestare contro il Governo ...

b) *PREP_NA*, for prepositions (e. g. "*del*", "*degli*" etc.) mainly governing a bare noun and yielding noun adjuncts;

... proporzione del vantaggio ...

c) *PREP_VA*, for prepositions (e. g. "*nella*", "*sul*" etc.) mainly governing a bare noun and forming verb adjuncts. The three prepositional patterns are exemplified below:

... provvedere in tempo ...

2.2. Data Preparation Notes

Each sentence in the data sets is considered a separate entity. The global amount of manually annotated data (slightly more than 151.000 tokens) has been split between DS and TS maintaining a ratio of 8/1. One sentence out of nine is extracted and inserted into TS. Following this schema we do not preserve text integrity, thus taggers cannot rely on it but will have to process each sentence separately.

3. TOKENISATION ISSUES

The problem of text segmentation (tokenisation) is a central issue in POS-tagger evaluation and comparison. In principle every system should apply different tokenisation rules leading to different outputs. In this first evaluation campaign we do not have the possibility of handling different tokenisation schemas and following the complex realignment work proposed, for example, inside the GRACE evaluation project [Adda *et al*. 1998].

In this EVALITA task we provide all the development and test data in tokenised format, one token per line followed by its tag (when applicable), following the schema:

<TOKEN_1> <TAG1>
<TOKEN_2> <TAG2>
...
<TOKEN_N> <TAGN>

Example:

```
Il              ART
dott.           NN
Rossi           NN_P
manger&agrave;  V_GVRB
le              ART
mele            NN
verdi           ADJ
dell'           PREP_A
orto            NN
```

```
di              PREP
Carlo           NN_P
fino_a          PREP
Natale          NN_P
.               P_EOS
```

The example above shows some tokenisation and formatting issues:

- accents are coded using ISO-Latin1 SGML entities (*mangerà*);
- the tokenisation process identified and managed abbreviations (*dott.*). The file `abbrev.txt` contains all the abbreviations considered during the process.
- apostrophe is tokenised separately only when used as quotation mark, not when signalling a removed character (dell'/orto);
- a list of multi-word expressions (MWE) has been considered: annotating MWE can be very difficult in some cases as we try to label them token-by-token, especially for expressions belonging to closed (grammatical) classes. Thus we decided to tokenise a list of these expressions as single units and to annotate them with a unique tag. The file `MWE.txt` contains the expressions we have tokenised in that way.

The participants are requested to return the test file using the same tokenisation format, containing exactly the same number of tokens. The comparison with the reference file will be performed line-by-line, thus a misalignment will produce wrong results.

The TS will not contain the correct tags; the correct results will be provided to the participants after the evaluation, together with their score.

4. EVALUATION METRICS

The evaluation is performed in a "black box" approach: only the systems' 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 number of correct PoS tag assignment divided by the total number of tokens in TS.

b) *Unknown Words Tagging Accuracy*: it is defined as the *Tagging Accuracy* restricting the computation to unknown words. In this context for "unknown word" we mean 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.

A baseline algorithm (Most Frequent Tag assignment) and some well known PoS-taggers (TnT, Brill and possibly some others) will be used as reference for comparison purposes.

5. EVALUATION DETAILS

The 1st March the task organiser will send to the registered participants (who have signed the license agreement) these Guidelines and the Development Set of data in the format described in section 3 by email. All the data will be provided as plain text files in UNIX format, thus pay attention to newline character format.

The 20th May the organiser will send the Test Set of data (tokenised, 1 token per line) by email; participants are required to return the tagged version of this file (without any change in the token stream) by the 1st June (midnight) naming the file as

`EVALITA07_POStask_participantname_TagSet` (with TagSet as "EAGLES" or "DISTRIB") and sending it to the organiser's email: *fabio.tamburini@unibo.it*. Only one version of this result file for each tagset will be accepted.

After the submission deadline the organiser will evaluate the systems' results and send back to the participants their score as well as the 'gold-standard' TS version.

REFERENCES

Adda, G., Mariani, J., Lecomte, J., Paroubek, P., Rajman, M. (1998), "The GRACE French Part-of-Speech Tagging Evaluation Task", in *Proceeedings of LREC'98*, Granada.

Bernardi, R., Bolognesi, A., Seidenari, C., Tamburini, F. (2005), "Automatic induction of a POS tagset for Italian". In *Proceeedings of ALTW 2005*, Sydney.

Bernardi, R., Bolognesi, A., Seidenari, C., Tamburini, F. (2006), "POS tagset design for Italian", in *Proceedings of LREC 2006*, Genova.

Bosco, C., Lombardo, V., Vassallo D., and Lesmo L. (2000), "Building a treebank for Italian: a data-driven annotation schema". In *Proceeedings of LREC 2000*, Athens.

Bosco, C. (2003), *A grammatical relation system for treebank annotation*. Ph.D. thesis, Computer Science Department, Turin University.

Dardano, M., Trifone, P. (1997), *La nuova grammatica della lingua italiana*, Bologna: Zanichelli.

De Mauro, T. (1999), *Grande dizionario dell'uso*, Torino: UTET.

De Mauro, T. (2007) Il dizionario della lingua italiana, On-line version http://www.demauroparavia.it/

Delmonte, R. (2004), "Strutture sintattiche dall'analisi computazionale di corpora di italiano". In A. Cardinaletti and F. Frasnedi (eds), *Intorno all'italiano contemporaneo. Tra linguistica e didattica*. Milano: F. Angeli.

Devoto, G., Oli, G. (2001), *Il dizionario della lingua italiana*, Firenze: Le Monnier.

Graffi, Giorgio 1994. *Le strutture del linguaggio. Sintassi*, Bologna: Il Mulino.

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

Renzi, Lorenzo / Salvi, Gianpaolo / Cardinaletti, Anna (eds.) 1988, 1991, 1995. *Grande grammatica italiana di consultazione*. Bologna: Il Mulino.

Sabatini, F., Coletti, V. (1997), *Disc: dizionario italiano Sabatini Coletti*. Firenze: Giunti.

Sensini, M. (1997), *Grammatica italiana*, Milano: Mondadori.

Serianni, L. (198), Grammatica italiana. Italiano comune e lingua letteraria , Torino: UTET.

Zingarelli, N. (2000), *Lo Zingarelli 2001: vocabolario della lingua italiana*, Bologna: Zanichelli