EVALITA 2011
THE ITALIAN LEMMATISATION EVALUATION
- TASK GUIDELINES -

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

1. INTRODUCTION

Lemmatisation, the process of transforming each word-form into its corresponding lemma, is often considered a subproduct of a part-of-speech procedure that does not cause any particular problem.

The common view is that no particular ambiguities have to be resolved once the correct PoS-tag has been assigned. Unfortunately there are a lot of specific cases, at least in Italian, in which, given the same lexical class, we face a lemma ambiguity.

The following table shows some examples:

WORDFORM	PoS-tag	POSSIBLE LEMMAS
cannone	NOUN	*cannone, canna*
morti	NOUN	*morto, morte*
regione	NOUN	*regione, regia*
aria	NOUN	*aria, ario*
macchina	NOUN	*macchina, macchia*
piccione	NOUN	*piccione, piccia*
matematica	NOUN	*matematica, matematico*
stazione	NOUN	*stazione, stazio*
passano	VERB	*passare, passire*
danno	VERB	*dare, dannare*

As you can see, homograph in verb forms belonging to different verbs or noun valutative suffixation are some phenomena that create such kind of lemma ambiguities.

Even the use of morphological analysers based on large lexica, which are undoubtedly very useful for the PoS-tagging procedures (see for example the results of the EVALITA2007 PoS-tagging task [Tamburini, 2007]), can create a lot of other ambiguities.

Certainly these phenomena are not pervasive and the total amount of such ambiguities is very limited, but we believe that it could be interesting to develop specific techniques to solve this generally underestimated problem.

The organisation will provide two data sets: the first, referred to as **Development Set (DS)** contains data manually classified (see a following section for a detailed description) and are to be used to set up participants' systems; the second, referred to as **Test Set (TS)** contains the test data for the evaluation. Lemmatisation is a complex process involving the entire lexicon. It is almost useless to provide a small set of training data for this task. No machine-learning algorithm will be able to acquire any useful information to solve this task using only some hundred thousand annotated tokens. Participants have to use or develop different kinds of approaches to face this task and they are allowed to use other resources in their systems, both for develop and to enhance final performances, but the results must conform to the proposed formats. The DS is provided only to check formats and specific decision about lemmatisation taken when developing the gold standard.

For the same reasons, we do not distribute a lexicon resource with EVALITA 2011 data. Each participant is allowed to use any available resource for Italian.

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 set used for this evaluation task is composed of the same data used in the EVALITA 2007 PoS-tagging task, considering the "EAGLES-like" tagset. These data have been manually annotated assigning to each token its lexical category (PoS-tag) and its correct lemma. Please refer to the EVALITA 2007 PoS-tagging task guidelines for a detailed description of the data set:
http://www.evalita.it/sites/evalita.fbk.eu/files/doc2007/EVALITA2007_POSTag_Guidelines.pdf

The organisation will provide the TS removing the lemma associated for each word form and each participant is required to apply its system and return the lemma assigned to each word form; only one solution for each token will be accepted.

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

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 1/8. One sentence out of nine is extracted and inserted into DS. Following this schema we do not preserve text integrity, thus applications 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 evaluation and comparison. In principle every system should apply different tokenisation rules leading to different outputs.

In this EVALITA task we provide all the test data in tokenised format, one token per line followed by its tag, 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.

During the evaluation, the comparison with the reference file will be performed line-by-line, thus a misalignment will produce wrong results.

The participants are requested to return the test file adding a third column containing exactly **one** lemma, using the same tokenisation format and the same number of tokens as in the following example:

```
Il              ART         Il
dott.           NN          dott.
Rossi           NN_P        Rossi
manger&agrave;  V_GVRB      mangiare
le              ART         le
mele            NN          mela
verdi           ADJ         verde
dell'           PREP_A      dell'
orto            NN          orto
di              PREP        di
Carlo           NN_P        Carlo
fino_a          PREP        fino_a
Natale          NN_P        Natale
.               P_EOS       .
```

The evaluation is only referred to open class words and not to functional words: only the tokens having a PoS-tag comprised in the set {`ADJ_*, ADV, NN, V_*`} have to be lemmatised, in all the other cases the token must be copied unchanged into the lemma column (the asterisk indicates all PoS-tag possibilities beginning with that prefix).

In case the token presents an apocope (*signor*, *poter*, *dormir*, …) the corresponding lemma has to be completed (*signore*, *potere*, *dormire*).

For cliticised verb forms (*mangiarlo*, *colpiscilo*, …), all the pronouns must be removed and the lemma must be the infinite verb form (*mangiare*, *colpire*).

The TS will not contain the correct lemmas; the gold standard 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 lemma is allowed for each token.

For this task we consider only one metric, the *lemmatisation accuracy*, defined as the number of correct lemma assignment divided by the total number of tokens in the test set belonging to the considered lexical classes (`ADJ_*, ADV, NN, V_*`).

The organisation will provide the scoring program during the development stage.

5. EVALUATION DETAILS

The beginning of June these guidelines will be available on the EVALITA 2011 Web site and the task organisers will send to the registered participants the DS. All the data sets will be provided as plain text files in UNIX format, thus pay attention to newline character format.

The 4th October the organisers will send the test data (tokenised, 1 token per line) by email; participants are required to return the lemmatised version of this file (without any change in the token and PoS-tag streams) by the 14th October (midnight) naming the file as `EVALITA11_LemmTask_participantname` and sending it to the organiser's email: *fabio.tamburini@unibo.it*. Only one version of this result file will be accepted.

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

Any information about the proceedings will appear to the EVALITA 2011 Web site (http://www.evalita.it/2011/proceedings).

REFERENCES

Tamburini F. (2007). EVALITA 2007: the Part-of-Speech Tagging Task. *Intelligenza Artificiale*, IV(2), 4-7, http://www.evalita.it/sites/evalita.fbk.eu/files/proceedings2007/01-IA-IV-2-pos-task.pdf .