The EVALITA 2011 Lemmatisation Task

Fabio Tamburini

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

Abstract. This paper reports on EVALITA 2011 Lemmatisation task, an initiative for the evaluation of automatic lemmatisation for Italian. A relevant number of scholars and teams participated experimenting their systems on the data provided by the task organisers. The results are very interesting and overall performances of the participating systems are very high.

Keywords: Evaluation, Lemmatisation, Italian

1 Motivation

Lemmatisation, the process of transforming each wordform into its corresponding base form found in the dictionary (lemma), is often considered a subproduct of a part-of-speech (PoS) 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:

Table 1. Examples of lemma ambiguities.

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*
osservatori	NOUN	*osservatore, osservatorio*
passano	VERB	*passare, passire*
danno	VERB	*dare, dannare*
perdono	VERB	*perdere, perdonare*

Homograph in verb forms belonging to different verbs or noun evaluative suffixation are some phenomena that can create such kind of lemma ambiguities.

There are lots of studies on the automatic learning of morphological rules able to connect each wordform to its respective lemma [1, 3], but there seems to be less interest in building automatic systems able to solve lemma ambiguities and assign the correct lemma "in context".

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 [4]), can create a lot of such ambiguities introducing more possibilities for creating homographs between different wordforms.

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.

2 Definition of the Task

The organisation provided two data sets: the first, referred to as Development Set (DS) contained a small set, composed of 17313 tokens, of data manually classified (see a following section for a detailed description) and were to be used to set up participants' systems; the second, referred to as Test Set (TS), contained the final test data for the evaluation and it was composed of 133756 tokens.

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 would be able to acquire any useful information to successfully solve this task using only some hundred thousand annotated tokens. For these reasons, participants had to use or develop different kinds of approaches to face this task; they were allowed to use other resources in their systems, both for develop and to enhance the final performances, but the results must be conformed to the proposed formats. The DS, then, was provided only to check formats and specific decisions about lemmatisation taken when developing the gold standard. For the same reasons, we did not distribute a lexicon resource with EVALITA 2011 data. Each participant was allowed to use any available resource for Italian. Participants were also required to send a brief description of the system, especially considering the techniques and resources used to develop their systems.

3 Dataset Description

The data set used for this evaluation task is composed of the same data used in the EVALITA 2007 Part-of-Speech 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. Table 2 shows the complete PoS-tagset used for this task.

The organisation provided the TS removing the lemma associated for each wordform and each participant was required to apply its system and return the lemma assigned to each wordform; only one solution for each token was accepted.

Table 2. EVALITA 2007 EAGLES-Like PoS-tagset used for this evaluation.

ADJ	Qualifying adjectives.	P_APO	Apostrophe as quotation mark.
ADJ_DIM	Demonstrative adjectives.	P_OTH	Other punctuation marks.
ADJ_IND	Indefinite adjectives.	PREP	Simple prepositions.
ADJ_IES	Interr. or excl. adjectives.	PREP_A	Prepositions fused with articles.
ADJ_POS	Possessive adjectives.	PRON_PER	Personal pronouns.
ADJ_NUM	Numeral adjectives.	PRON_REL	Relative pronouns.
ADV	Adverbs.	PRON_DIM	Demonstrative pronouns.
ART	Articles.	PRON_IND	Indefinite pronouns.
NN	Common nouns.	PRON_IES	Interrogative or exclamative pron.
NN_P	Proper Nouns.	PRON_POS	Possessive pronouns.
C_NUM	Cardinal numbers.	V_AVERE	All forms of *avere*.
CONJ_C	Coordinating conjunctions.	V_ESSERE	All forms of *essere*.
CONJ_S	Subordinating conjunctions.	V_MOD	All forms of *potere, dovere, volere*.
INT	Interjections.	V_PP	Past and present participles.
NULL	Symbols, codes, delimiters, ...	V_GVRB	General verb forms.
P_EOS	'.', '!', '?' closing a sentence.	V_CLIT	Cliticised verb forms (e.g. *andarci*).

3.1 Data Preparation Notes

Each sentence in the data sets was considered a separate entity. The global amount of manually annotated data (slightly more than 151000 tokens) has been split between DS and TS maintaining a ratio of 1/8. One sentence out of nine was extracted and inserted into DS. Following this schema we did not preserve text integrity; the various systems had to process each sentence separately.

3.2 Tokenisation Issues

The problem of text segmentation (tokenisation) is a central issue in evaluation and comparison. In principle every system could apply different tokenisation rules leading to different outputs. In this EVALITA task we provided all the test data in tokenised format, one token per line followed by its tag.

Example:

Token	PoS-tag	Lemma	Token	PoS-tag	Lemma
Il	ART	il	dell'	PREP_A	dell'
dott.	NN	dott.	orto	NN	orto
Rossi	NN_P	rossi	di	PREP	di
mangerà	V_GVRB	mangiare	Carlo	NN_P	carlo
le	ART	le	fino_a	PREP	fino_a
mele	NN	mela	Natale	NN_P	natale
verdi	ADJ	verde	.	P_EOS	.

The example above (that contains also the lemma column presenting the correct lemma for each token) shows some tokenisation and formatting issues:

– accents were coded using ISO-Latin1 SGML entities (*mangerà*) to avoid any problem of character set conversion;
– the tokenisation process identified and managed abbreviations (*dott.*). A list containing all the abbreviations considered during the process was provided to the participants.
– apostrophe was tokenised separately only when used as quotation mark, not when signalling a removed character (*dell'orto* → *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. Again, a list containing the expressions we have tokenised in this way was provided to the participants.

The participants were requested to return the test file adding a third column containing exactly one lemma, in lowercase format, using the same tokenisation format and the same number of tokens as in the example above. During the evaluation, the comparison with the gold standard was performed line-by-line, thus a misalignment produced wrong results.

4 Evaluation Procedures and Metrics

The evaluation was performed in a "black box" approach: only the systems' output was evaluated. The evaluation metrics were based on a token-by-token comparison and only one lemma was allowed for each token.

The evaluation was 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_* had to be lemmatised, in all the other cases the token could be copied unchanged into the lemma column as they were not considered for the evaluation (the asterisk indicates all PoS-tag possibilities beginning with that prefix). We chose to evaluate only tokens belonging to these classes because they represent the most interesting cases, the open classes. All the other lexical classes can be lemmatised in a straightforward way once decided the lemmatisation conventions for them.

In case the token presents an apocope (*signor*, *poter*, *dormir*, ...) the corresponding lemma had to be completed (*signore*, *potere*, *dormire*, ...). For cliticised verb forms (*mangiarlo*, *colpiscili*, ...), all the pronouns had to be removed and the lemma had to be the infinite verb form (*mangiare*, *colpire*, ...).

With regard to derivation, we did not require to convert the wordform to its base lemma except for evaluative suffixations and the suffix *-issimo* for superlatives.

The gold standard was provided to the participants after the evaluation, together with their score, to check their system output.

For this task we considered only one metric, the "Lemmatisation Accuracy", defined as the number of correct lemma assignments divided by the total number of tokens in the TS belonging to the lexical classes considered for the evaluation (65210 tokens). The organisation provided an official scoring program during the development stage in order to allow the participants to develop and evaluate their systems on the DS.

5 Participants and Results

Four systems participated to the final evaluation, three from Italy and one from France. Table 3 shows some details of the research groups that participate to the task.

Table 3. Lemmatisation Task participants.

Name	Institution	System Label
Rodolfo Delmonte	University of Venice, Italy	Delmonte_UniVE
Djamé Seddah	Alpage (Inria)/Univ. Paris Sorbonne, France	Seddah_Inria-UniSorbonne
Maria Simi	University of Pisa, Italy	Simi_UniPI
Fabio Tamburini	University of Bologna, Italy	Tamburini_UniBO

The structure of the participating systems is carefully described in specific papers contained in this proceedings volume. Here we would like to briefly sketch some of their basic properties and applied procedures:

- *Delmonte_UniVE* - a rule based lemmatiser based on a lexicon composed of about 80.000 lemmas and additional modules for managing ambiguities based on frequency information extracted from various sources.
- *Seddah_Inria-UniSorbonne* - a tool for supervised learning of inflectional morphology as a base for building a PoS-tagger and a lemmatiser and a lexicon extracted from Morph-It [6] and the Turin University Treebank [5].
- *Simi_UniPI* - a basic lemmatiser based on about 1.3 millions of wordforms followed by a cascade of filters (affix specific management, search in Wikipedia or directly on Google for similar contexts, ...).
- *Tamburini_UniBO* - a lemmatiser based on Finite State Automata equipped with a large lexicon of 110.000 lemmas and a simple algorithm that relies on the lemma frequency classification proposed in the De Mauro/Paravia dictionary [2].

Four, very simple and naïve, baseline systems were introduced by the organisers. The first system, *Baseline_1*, simply copied the input wordform into the output lemma. The second baseline, *Baseline_2*, acted as the first but corrected the output lemma for some simple cases:

- in case the PoS-tag was V_ESSERE or V_AVERE it replaced the lemma with, respectively, the verb infinitives *essere* or *avere*.
- in case the PoS-tag was V_MOD it replaced the output lemma with one of the infinitives *potere, volere, dovere* by simply looking at the first letter of the input wordform.

The third baseline, *Baseline_3*, followed the same procedure of Baseline_2 but, in case the two rules on PoS-tags did not apply, chose the lemma from the De Mauro/Paravia online dictionary [2] exhibiting the smallest Levenshtein distance with the examined wordform. The last baseline, *Baseline_4*, is a modification of Baseline_3: it searches into

the DS lexicon for a reference lemma before applying any heuristics on orthographic forms.

Table 4 outlines the results obtained by the various systems and by the baselines in terms of Lemmatisation Accuracy.

Table 4. EVALITA 2011 Lemmatisation Task results.

System	Lemmatisation Accuracy
Simi_UniPI	99.06%
Tamburini_UniBo	98.74%
Delmonte_UniVE	98.42%
Seddah_Inria-UniSorbonne	94.76%
Baseline_4	83.42%
Baseline_3	66.20%
Baseline_2	59.46%
Baseline_1	50.27%

6 Discussion

In this section we will try to draw some provisional conclusions about this task.

The results obtained by the participating systems were quite hight, mostly of them above 98% of Lemmatisation Accuracy. Considering that only half of the total number of tokens in the TS have been evaluated, these results depict a good global picture for this evaluation task. We can say that most of the ambiguities found in the test corpus were successfully solved by the most performant systems.

The neat separation between the baselines performances and the real systems can suggest that this task cannot be solved by using simple heuristics, but the disambiguation process has to be based on various sources of information: large lexica, frequency lists, powerful lemmatiser morphology-aware and so on. *Baseline_4*, the unique baseline using a lexicon of correct classifications, performs much better than the other baselines, but its performance is still not comparable with real systems.

Only the best performing system, in our knowledge, use the sentence context to choose among the different lemmas connected to an ambiguous wordform. Maybe this could be the most promising direction for increasing the automatic system performances for the lemmatisation task.

References

1. Creutz, M., Lagus, K.: Unsupervised models for morpheme segmentation and morphology learning. ACM Transactions on Speech and Language Processing, 4(1), pp. 3:1–3:34 (2007)
2. De Mauro, T.: Il dizionario della lingua italiana, Paravia (2000)

3. Hammarström, H., Borin, L.: Unsupervised Learning of Morphology. Computational Linguistics, 37(2), pp. 309–350 (2011)
4. Tamburini, F.: EVALITA 2007: the Part-of-Speech Tagging Task. Intelligenza Artificiale IV(2), 4–7 (2007)
5. The Turing University Treebank. http://www.di.unito.it/∼tutreeb
6. Zanchetta E., Baroni, M.: Morph-it! A free corpus-based morphological resource for the Italian language. In: Proceedings of Corpus Linguistics 2005, University of Birmingham (2005)