The Lemmatisation Task at the EVALITA 2011 Evaluation Campaign

Fabio Tamburini

Dept. of Linguistics and Oriental Studies, University of Bologna, Italy
fabio.tamburini@unibo.it

Abstract. This paper reports on the EVALITA 2011 Lemmatisation task, an initiative for the evaluation of automatic lemmatisation tools specifically developed for the Italian language. Despite lemmatisation is often considered a subproduct of a PoS-tagging procedure that does not cause any particular problem, there are a lot of specific cases, certainly in Italian and in some other highly inflected languages, in which, given the same lexical class, we face a lemma ambiguity. 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 the overall performances of the participating systems were very high, exceeding, on interesting cases, 99% of lemmatisation accuracy.

Keywords: Evaluation, Lemmatisation, Italian.

1 Introduction

In the general linguistics view, lemmatisation is the process of grouping together the different inflected forms of a word so they can be analysed as a single item[1].

In computational linguistics, usually, two different methods are used to achieve this task: the first, called *stemming*, tries to reduce all the wordforms belonging to a specific paradigm to an invariant stem string, by removing all affixes, and does not, in general, produce a real surface string. The second, *lemmatisation*, identifies the process of transforming each wordform into its corresponding canonical base form, the lemma, where the canonical form is one particular wordform from an inflectional paradigm chosen by convention to represent the whole paradigm and, usually, corresponds to a headword found in a dictionary. In Italian, canonical base forms corresponds to verb inifinitives and the masculine singular form for nouns and adjectives (except for those cases that allow only the feminine gender).

Lemmatisation and stemming are normalisation techniques which proved to be very useful in a number of different NLP tasks, for information extraction and retrieval and to simplify corpus querying. The use of such normalisation methods helps automatic retrieval systems to remove wordform differences due to inflectional phenomena. They are both very relevant for highly inflected languages, for example romance, slavic and some northern european languages as well as a lot of other languages around the world, where the co-selection between bases and the different kind of affixes, both inflectional

[1] Collins English Dictionary, entry for "lemmatise".

B. Magnini et al. (Eds.): EVALITA 2012, LNCS 7689, pp. 230–238, 2013.

and derivational, can depend on a whole range of factors, from phonological to semantic (see [10] for a description of the different degree of inflection across languages).

In homograph handling we face essentially two types of ambiguities: *internal or grammatical ambiguities* when we encounter different wordforms belonging to the same lemma and consequently to the same part-of-speech (PoS) tag (e.g. *ami* as different forms of the verb *amare* - to love), and *external or lexical ambiguities* when considering wordforms belonging to different lemmas, but not necessarily to different PoS-tags (e.g. the verb form *perdono* in Table 1). Internal ambiguities do not matter for the lemmatisation task, because we should assign the same lemma, but for external ambiguities we face two very different cases: the first involves different PoS-tags and this is sufficient for choosing the correct lemma, but in the second case we can have two different lemmas presenting the same PoS-tag.

In the current literature, lemmatisation is often considered a subproduct of a PoS-tagging 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 and a lot of the systems handling this task for different languages assume this view without indentifying and discussing the remaining potential external ambiguities [1,2,6,8,11,14], while some other scholars recognise the potential problem but ignore it [7].

Unfortunately there are a lot of specific cases, certainly in Italian and in some other highly inflected languages, in which, given the same lexical class, we face an external lemma ambiguity. The Table 1 shows some examples of such ambiguities for Italian. Homograph in verb forms belonging to different verbs or noun evaluative suffixation and plural forms are some phenomena that can create such kind of lemma ambiguities. A morphologically richer PoS-tagset could help alleviating the problem, at the price of a reduction in tagging accuracy, but in some cases the lemma ambiguity still persists.

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 [12]), 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 the 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

Table 1. Some examples of external lemma ambiguities in Italian

Wordform	Word class	Possible Lemmas
cannone	NOUN	*cannone, canna*
regione	NOUN	*regione, regia*
stazione	NOUN	*stazione, stazio*
piccione	NOUN	*piccione, piccia*
macchina	NOUN	*macchina, macchia*
aria	NOUN	*aria, ario*
matematica	NOUN	*matematica, matematico*
osservatori	NOUN	*osservatore, osservatorio*
morti	NOUN	*morto, morte*
cocchi	NOUN	*cocchio, cocco*
acerrima	ADJ	*acerrimo, acre*
molli	ADJ	*molle, mollo*
nubili	ADJ	*nubile, nubilo*
sanzionatori	ADJ	*sanzionatori, sanzionatorio*
butterò	VERB	*butterare, buttare*
danno	VERB	*dare, dannare*
dura	VERB	*durire, durare*
fonda	VERB	*fondare, fondere*
incappi	VERB	*incappare, incappiare*
passano	VERB	*passare, passire*
perdono	VERB	*perdere, perdonare*
smaltiate	VERB	*smaltare, smaltire*
stecchiate	VERB	*stecchire, steccare*
veniamo	VERB	*venire, venare*

use or develop different kinds of approaches to face this task; they were allowed to use other resources in their systems, both to 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.

The proposed tagset is designed taking as reference the EAGLES guidelines [9]. 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 [4].

Table 2 shows the complete PoS-tagset used for this task.

Table 2. EVALITA 2007 EAGLES-Like PoS-tagset used for this Lemmatisation-Task 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*).

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. We decided to annotate as NN_Ps 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

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.

All the available data have been manually annotated assigning to each token its lexical category (PoS-tag) and its correct lemma. 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.

3.1 Data Preparation Notes

Each sentence in the data sets was considered as 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, thus 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
Il	ART	il
dott.	NN	dott.
Rossi	NN_P	rossi
mangerà	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 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* $\rightarrow$ *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.

The structure of the participating systems is carefully described in specific papers contained in this 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 roots and additional modules for managing ambiguities based on frequency information extracted from various sources.

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

- *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 [15] and the Turin University Treebank [13].
- *Simi_UniPI* - an independent PoS-tagger with 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 derived from a Morphological Analyser based on Finite State Automata and 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 [4].

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 (as in [1]). 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 character 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 [4] 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 official results obtained by the various systems and by the baselines in terms of Lemmatisation Accuracy.

In tables 5 and 6 we made some analysis of the errors produced by the participating systems. The first table presents the distribution of the errors between the four different lexical classes considered in the evaluation, computed dividing the system error in a specific class by the total number of errors made by the system. The other table analyses the errors inside each specific class and measure the amount of errors made by the system dividing them by the total number of tokens belonging to the same class in TS.

Considering the best three performing systems, we can note that most of their errors are concentrated on nouns: annotating the NN PoS-class, they exhibits the highest error rate both considering the absolute picture and considering the relative intra-class error.

Table 4. EVALITA 2011 Lemmatisation Task official 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%

Table 5. Systems' absolute error distribution with respect to PoS-tags (computed as the error for each class divided by the total number of errors made by the system)

System	ADJ_*	ADV	NN	V_*
Simi_UniPI	15.6%	8.2%	61.2%	15.0%
Tamburini_UniBo	17.7%	5.1%	64.4%	12.8%
Delmonte_UniVE	11.9%	6.7%	70.8%	10.6%
Seddah_Inria-UniSorbonne	25.6%	4.9%	30.4%	44.1%

Table 6. Systems' relative error inside each lexical class (computed as the error made by the system for each class divided by the total number of token in the same class contained into the TS)

System	ADJ_*	ADV	NN	V_*
Simi_UniPI	0.8%	0.7%	1.4%	0.5%
Tamburini_UniBo	1.2%	0.6%	2.0%	0.5%
Delmonte_UniVE	1.0%	1.0%	2.7%	0.6%
Seddah_Inria-UniSorbonne	7.0%	2.4%	3.9%	8.1%

One possible explanation concerns the high complexity of the evaluative morphology in Italian that is able to create a lot of potential homograph for nouns and adjectives. This consideration can be further supported by noting that the adjective class is the second problematic category for these systems.

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, and that the other half should not create any problem at all, 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, not surprisingly, the most promising direction for increasing the automatic system performances for the lemmatisation task. The same system applied a different PoS-tagger to enrich the morphological information available to the lemmatiser for disambiguating lemma ambiguities: this could be, as we argued before, a viable solution to reduce the number of real ambiguity cases, but it has to be carefully balanced with the unavoidable reduction in performance of the PoS-tagger.

References

1. Agic, Z., Tadic, M., Dovedan, Z.: Evaluating Full Lemmatization of Croatian Texts. Recent Advances in Intelligent Information Systems, pp. 175–184. Academic Publishing House (2009)
2. Airio, E.: Word normalization and decompounding in mono- and bilingual. IR Information Retrieval 9, 249–271 (2006)
3. Creutz, M., Lagus, K.: Unsupervised models for morpheme segmentation and morphology learning. ACM Transactions on Speech and Language Processing 4(1), 3:1–3:34 (2007)
4. De Mauro, T.: Il dizionario della lingua italiana, Paravia (2000)
5. Hammarström, H., Borin, L.: Unsupervised Learning of Morphology. Computational Linguistics 37(2), 309–350 (2011)
6. Hardie, A., Lohani Yogendra, R.R., Yadava, P.: Extending corpus annotation of Nepali: advances in tokenisation and lemmatisation. Himalayan Linguistics 10(1), 151–165 (2011)
7. Ingason, A.K., Helgadóttir, S., Loftsson, H., Rögnvaldsson, E.: A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI). In: Nordström, B., Ranta, A. (eds.) GoTAL 2008. LNCS (LNAI), vol. 5221, pp. 205–216. Springer, Heidelberg (2008)
8. Mendes, A., Amaro, R., Bacelar do Nascimento, M.F.: Reusing Available Resources for Tagging a Spoken Portuguese Corpus. In: Branco, A., Mendes, A., Ribeiro, R. (eds.) Language Technology for Portuguese: Shallow Processing Tools and Resources, pp. 25–28. Lisbon, Edicoes Colibri (2003)
9. Monachini, M.: ELM-IT: EAGLES Specification for Italian morphosintax Lexicon Specification and Classification Guidelines. EAGLES Document EAG CLWG ELM IT/F (1996)
10. Pirkola, A.: Morphological typology of languages for IR. Journal of Documentation 57(3), 330–348 (2001)
11. Plisson, J., Lavrač, N., Mladenić, D., Erjavec, T.: Ripple Down Rule Learning for Automated Word Lemmatisation. AI Communications 21, 15–26 (2008)
12. Tamburini, F.: EVALITA 2007: the Part-of-Speech Tagging Task. Intelligenza Artificiale IV(2), 4–7 (2007)
13. The Turin University Treebank, `http://www.di.unito.it/~tutreeb`
14. Van Eynde, F., Zavrel, J., Daelemans, W.: Lemmatisation and morphosyntactic annotation for the spoken Dutch corpus. In: Proceedings of CLIN 1999, pp. 53–62. Utrecht Institute of Linguistics OTS, Utrecht (1999)
15. 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)