The AnIta-Lemmatiser: A Tool for Accurate Lemmatisation of Italian Texts

Fabio Tamburini

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

Abstract. This paper presents the AnIta-Lemmatiser, an automatic tool to lemmatise Italian texts. It is based on a powerful morphological analyser enriched with a large lexicon and some heuristic techniques to select the most appropriate lemma among those that can be morphologically associated to an ambiguous wordform. The heuristics are essentially based on the frequency-of-use tags provided by the De Mauro/Paravia electronic dictionary. The AnIta-Lemmatiser ranked at the second place in the Lemmatisation Task of the EVALITA 2011 evaluation campaign. Beyond the official lemmatiser used for EVALITA, some further improvements are presented.

Keywords: Lemmatisation, Italian, Morphological Analyser, Lexicon.

1 Introduction

Stemming and lemmatisation are fundamental normalisation tasks at low-level Natural Language Processing (NLP), in particular for morphologically complex languages involving rich inflectional and derivational phenomena.

In the current literature, lemmatisation 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 and a lot of the systems handling this task for different languages assume this view without indentifying and discussing the remaining potential external ambiguities [1,2,14,19,21,28], while some other scholars recognise the potential problem but ignore it [15].

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. Thus, a successful lemmatiser has to implement specific techniques to deal with these ambiguities in a proper way.

Current, state-of-the-art lemmatisers are usually based on powerful morphological analysers able to handle the complex information and processes involved in successful wordform analysis.

The system described in this paper has been developed to lemmatise CORIS, a large reference corpus of contemporary written Italian [23]; it is based on a powerful morphological analyser enriched with a large lexicon and some heuristic techniques to select the most appropriate lemma among those that can be morphologically associated to an ambiguous wordform.

B. Magnini et al. (Eds.): EVALITA 2012, LNCS 7689, pp. 266–273, 2013.

After the seminal work of Koskenniemi [17] (see also the recent books [5,22] for general overviews) introducing the two-level approach to computational morphology, a lot of successful implementations of morphological analysers for different languages has been produced [5,7,20,24,27]. Although this model has been heavily challenged by some languages (especially semitic languages [12,16]), it is still the reference model for building such kind of computational resources.

In the late nineties some corpus-based/machine-learning methods were introduced to automatically induce the information for building a morphological analyser from corpus texts (see the review papers [8,13]). These methods seem to be able to induce the lexicon from data, avoiding the complex work of manually writing it, despite some reduction in performance.

Italian is one of the ten most widely spoken languages in the world. It is a highly-inflected Romance language: words belonging to inflected classes (adjectives, nouns, determiners and verbs) exhibit a rich set of inflection phenomena. Noun inflection, also shared with adjectives and determiners, has different suffixes for gender and number, while verb inflection presents a rich set of regular inflections and a wide range of irregular behaviours. All inflection phenomena are realised by using different suffixes. Nouns, adjectives and verbs form the base for deriving new words through complex combinations of prefixes and suffixes. Also compounded forms are quite frequent in Italian.

From a computational point of view there are some resources able to manage the complex morphological information of the Italian language. On the one hand we have open source or freely available resources, such as:

- *Morph-it* [30] an open source lexicon that can be compiled using various packages implementing Finite State Automata (FSA) for two-level morphology (SFST-Stuttgart Finite State Transducer Tools and Jan Daciuk's FSA utilities). It globally contains 505,074 wordforms and 35,056 lemmas. The lexicon is quite small and, in order to be used to successfully annotate real texts, it requires to be extended. Moreover, the lexicon is presented as an annotated wordform list and extending it is a very complex task. Although it uses FSA packages it does not exploit the possibilities provided by these models of combining bases with inflection suffixes, thus the addition of new lemmas and wordforms requires listing all possible cases.
- *TextPro/MorphoPro* [20] a freely available package (only for research purposes) implementing various low-level and middle-level tasks useful for NLP. The lexicon used by MorphoPro is composed of about 89,000 lemmas, but, being inserted into a closed system, it cannot be extended in any way. The underlying model is based on FSA.

On the other side we have some tools not freely distributed that implement powerful morphological analysers for Italian:

- *MAGIC* [4] is a complex platform to analyse and generate Italian wordforms based on a lexicon composed of about 100,000 lemmas. The lexicon is quite large, but it is not available to the research community; ALEP is the underlying formalism used by this resource.
- *Getarun* [9] is a complete package for text analysis. It contains a wide variety of specific tools to perform various NLP tasks (PoS-tagging, parsing, lemmatisation,

anaphora resolution, semantic interpretation, discourse modelling...). Specifically, the morphological analyser is based on 80,000 roots and large lists of about 100,000 wordforms. Again the lexicon is quite large, but, being a closed application not available to the community, it does not allow to profitably use such resource to develop new NLP tools for the Italian language.

1.1 The AnIta Morphological Analiser

This section briefly describes *AnIta*, a morphological analyser for Italian based on a large hand-written lexicon and two-level, rule-based finite-state technologies (for a detailed description see [26]).

The motivations for the choice of such model can be traced back, on the one hand, to the availability of a large electronic lexicon ready to be converted for these models and, on the other hand, on the the aim of obtaining an extremely precise and performant tool able to cover a large part of the wordforms found into real Italian texts (this second requirement drove us to choose a rule-based manually-written system, instead of unsupervised machine-learning methods, for designing the lexicon).

It is quite common, in computational analysis of morphology, to implement systems covering most of the inflectional phenomena involved in the studied language. Implementing the management of derivational and compositional phenomena in the same computational environment is less common and morphological analysers covering such operations are quite rare (e.g. [24,27]).

The implementation of derivational phenomena in Italian, considering the framework of two-level morphology, has been extensively studied by [6]; the author concludes that "...the continuation classes representing the mutual ordering of the affixes in the word structure are not powerful enough to provide a motivated account of the co-selectional restriction constraining affixal combination. In fact, affix co-selection is sensitive to semantic properties." Considering this results we decided to implement only the inflectional phenomena of Italian by using the considered framework and manage the other morphological operations by means of a different annotation scheme.

The development of the AnIta morphological analyser is based on the Helsinki Finite-State Transducer package [18].

Considering the morphotactics combinations allowed for Italian, we have currently defined about 110,000 lemmas, 21,000 of which without inflection, 51 continuation classes (or inflectional classes) to handle regular and irregular verb conjugations (following the proposal of [3] for the latter) and 54 continuation classes for noun and adjective declensions. In Italian clitic pronouns can be attached to the end of some verbal forms and can be combined together to build complex clitic clusters. All these phenomena have been managed by the analyser through specific continuation classes.

Nine morphographemic rules handle the transformations between abstract lexical strings and surface strings, mainly for managing the presence of velar and glide sound in the edge between the base and the inflectional suffix. An example of such a rule is the cancellation of the last letter '*i*' in the base if the inflectional suffix begins with a '*ia*' diphthong (e.g. the verb *marciare* - to march - will become *marc+iamo* and not *marci+iamo* at the first person plural of present indicative).

We also added 3,461 proper nouns from person names, countries, cities and Italian politicians surnames to the AnIta lexicon in order to increase the coverage for this word class in real texts.

Table 1 shows some examples of AnIta morphological analyses.

Table 1. Some examples of AnIta analyses provided by the Morphological Analyser

Wordform	Morphological analysis
adulti	l_adulto+NN+MASC+PLUR
	l_adulto+ADJ+MASC+PLUR
ricercai	l_ricercare+V_FIN+IND+PAST+1+SING
mangiarglielo	l_mangiare+V_NOFIN+INF+PRES+C_GLI+C_LO
impareggiabile	l_impareggiabile+ADJ+FEMM+SING
capostazione	l_capostazione+NN+MASC+SING

1.2 The AnIta Lemmatiser

As outlined before, the availability of a large morphological analyser for Italian became fundamental for developing a performant lemmatiser; the AnIta lexicon contains a very large quantity of Italian lemmas and it is able to generate and recognise millions of wordforms and assign them to a proper lemma (or lemmas). Testing the analyser coverage on CORIS, we found that 97.21% of corpus tokens were recognised. For testing, we considered only wordforms satisfying the regular expression `/[a-zA-Z]+'?/`, as the purpose of this evaluation was to test the analyser on real words excluding all non-words (numbers, codes, acronyms, ...), quite frequent in real texts [26].

Unfortunately, the morphological analyser cannot disambiguate the cases in which the wordform is ambiguous both from an orthographic and grammatical point of view (see [25] for some examples). For this reason we have to introduce specific techniques to post-process the morphological analyser output when we encounter a lemma ambiguity.

The lemmatisation task can hardly be faced by using techniques that rely on machine learning processes because, in general, we do not have enough manually annotated data to successfully train such models and, in particular, the Development Corpus provided by the organisers was very small. A successful disambiguation process based on learning methods would require several millions of wordforms manually annotated with the correct lemma in order to be able to capture the subtle distinctions of the various lemmas.

The AnIta lemmatiser uses a very simple technique: in case of ambiguity between two or more lemmas the lemmatiser choose the most frequent one, but estimating the lemma frequency without a large lemmatised corpus is, indeed, a very complex task. We decided to use the estimation proposed by De Mauro in his pioneering work [10] when applied to the De Mauro/Paravia online dictionary [11]. This dictionary contains, for each sense of every lemma, a specific annotation that represents a mix of the lemma frequency and its dispersion across different text genres. Using these annotations (see Table 2) we can simply assign to every ambiguous wordform the most frequent lemma by considering the sorting depicted in the table.

Table 2. Frequency-of-use tags in the De Mauro/Paravia dictionary

1) FO *Fondamentale* - Fundamental	7) RE *Regionale* - Regional
2) AU *Alto uso* - High use	8) DI *Dialettale* - Dialectal
3) AD *Alta disponibilità* - High availability	9) ES *Esotismo* - Esotic
4) CO *Comune* - Common	10) BU *Basso uso* - Low use
5) TS *Tecnico/specialistico* - Technical	11) OB *Obsoleto* - Obsolete
6) LE *Letterario* - Literary	

This lemma classification is quite broad and a lot of different lemmas (more than 10,000 very frequent lemmas) are classified in the first three classes. In the next section we will discuss this problem in detail and propose some viable solutions.

2 Results and Discussion

Table 3 shows the lemmatisation task official results for the EVALITA 2011 evaluation campaign: the AnIta Lemmatiser, even using a simple frequency based technique for disambiguating among the possible lemmas associated to an ambiguous wordform, produced accurate results arriving at the second place in the official global evaluation ranking.

In order to quantify the improvement of the heuristic based on the De Mauro frequency classification extracted from his dictionary, we tested also a different version of our system that randomly chooses one of the possible lemmas associated, by the AnIta morphological analyser, to an ambiguous wordform. This "baseline"-AnIta-based system (AnIta-Random) is less performant, confirming that the frequency-based heuristic is able to produce appreciable improvements.

After the end of the evaluation, we produced a new version of the AnIta-Lemmatiser that uses the Development Set (DS) lexicon to increase the performances. We have to note that the classification of the De Mauro dictionary are quite broad and it is not infrequent that some of the ambiguous lemmas connected to a specific wordform lies in the same frequency class. Adapting the behaviour of the lemmatiser to the specific text type by applying the information extracted from the DS lexicon to the frequency-based selection procedure, improved the results (*AnIta-Lemmatiser-Improved*).

In order to identify the weakness of the AnIta Lemmatiser, it is worth to analise the kind of errors produced by the proposed system. Table 4 shows two different error analyses: the first line depicts the system 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; the second line shows the system relative error inside each lexical class, computed as the error made for each class divided by the total number of token in the same class contained into the Test Set (TS).

Most of the lemmatiser errors are concentrated on nouns: annotating the NN PoS-class, it exhibits the highest error rate both considering the absolute picture (64.4%) and considering the relative intra-class error (2.0%). 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

Table 3. EVALITA 2011 Lemmatisation Task official results

System	Lemmatisation Accuracy
1st Participant	99.06%
AnIta-Lemmatiser-WSM	98.92%
AnIta-Lemmatiser-Improved	98.87%
AnIta-Lemmatiser	98.74%
3rd Participant	98.42%
AnIta-Random	97.19%
4th Participant	94.76%
Baseline_4	83.42%
Baseline_3	66.20%
Baseline_2	59.46%
Baseline_1	50.27%

Table 4. System error analysis

System	ADJ_*	ADV	NN	V_*
Absolute error distribution with respect to PoS-tags.	17.7%	5.1%	64.4%	12.8%
Relative intra-class error inside each lexical class.	1.2%	0.6%	2.0%	0.5%

supported by noting that the adjective class is the second problematic category for the AnIta Lemmatiser.

A lot of further improvements can be introduced considering the information provided by the immediate context of the ambiguous wordform: agreement tests, the introduction of a light semantic information processing, for example by using a Word Space Model (WSM) of the sentence, and a refined frequency classification can be considered viable techniques to improve the overall performance of the AnIta Lemmatiser.

A specific improvement we are currently testing concerns the use of a WSM as a source of contextual information. We can introduce a simple bayesian model for choosing the correct lemma in case of ambiguity:

$$\overline{l(w)} = \underset{l_i(w)}{argmax}\ P(l_i(w)|C) = \underset{l_i(w)}{argmax}\ P(C|l_i(w)) \cdot P(l_i(w)) \tag{1}$$

where $l_1, l_2, ..., l_n$ are the various lemmas that can be associated to the ambiguous word w and C represents the context of w (for example, the sentence containing w). We can estimate the two probabilities involved in this model in different ways. For example we can use a WSM for estimating the dependence of the considered context C from each possible lemma l_i, $P(C|l_i(w))$, by using the cosine similarity provided by a WSM as a probability approximation and the frequency estimation derived by the De Mauro dictionary, as explained before, as an estimate of the lemma probability $P(l_i(w))$.

We have slightly modified the AnIta-Lemmatiser (starting from the 'Improved' version presented before) to apply this distributional model only in case of lemma ambiguities that cannot be resolved with the frequency-based algorithm: these cases occur when two or more lemmas have identical frequency estimations, a situation quite common given the rough estimations provided by the De Mauro dictionary through the

classes showed in table 2. The WSM was built using CORIS and the InfoMap-NLP package [29].

This newer version of the AnIta-Lemmatiser (*AnIta-Lemmatiser-WSM*) obtained a slight improvement when evaluated on the EVALITA Test Set, scoring 98.92% of Lemmatisation Accuracy.

Using such kind of models allow us to take into consideration the context in which the word we have to disambiguate lies. As discussed at the end of the evaluation [25], the use of the context seems to be, not surprisingly, one of the most promising source of information also for lemmatisation systems.

Currently, we are testing the Lemmatiser on the annotation of the CORIS/CODIS corpus and the results are, qualitatively, quite satisfactory.

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. Battista, M., Pirrelli, V.: Monotonic Paradigmatic Schemata in Italian Verb Inflexion. In: Proc. of COLING 1996, Copenhagen, pp. 77–82 (1996)
4. Battista, M., Pirrelli, V.: Una piattaforma di morfologia computazionale per l'analisi e la generazione delle parole italiane. ILC-CNR (2000)
5. Beesley, K.R., Karttunen, L.: Finite State Morphology. CSLI Publications (2003)
6. Carota, F.: Derivational Morphology of Italian: Principles for Formalisation. Literary and Linguistic Computing 21, 41–53 (2006)
7. Cöltekin, C.: A Freely Available Morphological Analyzer for Turkish. In: Proc. of the 7th International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta (2010)
8. 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)
9. Delmonte, R.: Computational Linguistic Text Processing - Lexicon, Grammar, Parsing and Anaphora Resolution. Nova Science Publisher, New York (2009)
10. De Mauro, T.: Guida all'uso delle parole. Editori Riuniti, Roma (1980)
11. De Mauro, T.: Il dizionario della lingua italiana, Paravia (2000)
12. Gridach, M., Chenfour, N.: XMODEL: An XML-based Morphological Analyzer for Arabic Language. International Journal of Computational Linguistics 1(2), 12–26 (2010)
13. Hammarström, H., Borin, L.: Unsupervised Learning of Morphology. Computational Linguistics 37(2), 309–350 (2011)
14. 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)
15. 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)
16. Kiraz, G.A.: Computational Nonlinear Morphology: with emphasis on Semitic Languages. Cambridge University Press (2004)
17. Koskenniemi, K.: Two-level morphology: A general computational model for word-form recognition and generation. PhD Thesis, University of Helsinki (1983)

18. Lindén, K., Silfverberg, M., Pirinen, T.: HFST Tools for Morphology - An Efficient Open-Source Package for Construction of Morphological Analyzers. In: Proc. of the Workshop on Systems and Frameworks for Computational Morphology, Zurich (2009)
19. 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, Lisbon, Edicoes Colibri, pp. 25–28 (2003)
20. Pianta, E., Girardi, C., Zanoli, R.: The TextPro tool suite. In: Proc. of the 6th Language Resources and Evaluation Conference (LREC 2008), Marrakech (2008)
21. Plisson, J., Lavrač, N., Mladenić, D., Erjavec, T.: Ripple Down Rule Learning for Automated Word Lemmatisation. AI Communications 21, 15–26 (2008)
22. Roark, B., Sproat, R.: Computational Approaches to Morphology and Syntax. Oxford University Press (2006)
23. Rossini Favretti, R., Tamburini, F., De Santis, C.: CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model. In: Wilson, A., Rayson, P., McEnery, T. (eds.) A Rainbow of Corpora: Corpus Linguistics and the Languages of the World, Lincom-Europa, Munich, pp. 27–38 (2002)
24. Schmid, H., Fitschen, A., Heid, U.: SMOR: A German computational morphology covering derivation, composition, and inflection. In: Proc. of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, pp. 1263–1266 (2004)
25. Tamburini, F.: The EVALITA 2011 Lemmatisation Task. In: Working Notes of EVALITA 2011, Rome, Italy (January 24-25, 2012)
26. Tamburini, F., Melandri, M.: AnIta: a powerful morphological analyser for Italian. In: Proc. of LREC 2012, Istanbul, pp. 941–947 (2012)
27. Tzoukermann, E., Libermann, M.Y.: A finite-state morphological processor for Spanish. In: Proc. of COLING 1990, pp. 277–281 (1990)
28. 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)
29. Widdows, D.: Geometry and Meaning. CSLI Publication (2004)
30. Zanchetta, E., Baroni, M.: Morph-it! A free corpus-based morphological resource for the Italian language. In: Proc. Corpus Linguistics 2005, Birmingham (2005)