The AnIta-Lemmatiser

Fabio Tamburini

Dipartimento di Studi Linguistici e Orientali, Università di 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.

Keywords: Lemmatisation, Italian, Morphological Analyser, Lexicon

1 Description of the System

Stemming and lemmatisation are fundamental tasks at low-level Natural Language Processing (NLP) in particular for morphologically complex languages involving rich inflectional and derivational phenomena. These tasks are usually based on powerful morphological analysers able to handle the complex information and processes involved in successful wordform analysis.

After the seminal work of Koskenniemi [13] (see also the recent books [3, 16] for general overviews) introducing the two-level approach to computational morphology, a lot of successful implementations of morphological analysers for Western European languages has been produced [3, 5, 15, 18, 20]. Although this model has been heavily challenged by some languages (especially semitic languages [10, 12]), it is still the reference model for building such kind of computational resources at least for Western European languages.

In the late nineties some corpus-based/machine-learning methods were introduced to automatically induce the information for building a morphological analyser from corpus data (see the review papers [6, 11]). 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, pronouns and verbs) exhibit a rich set of inflection phenomena. Noun inflection, also shared with adjectives, determiners and pronouns, 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* [21] 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* [15] 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* [2] 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* [7] 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 lemmas and large lists of about 100,000 wordforms. Again the lexicon is quite large, but, being a close 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 AnIta Morphological Analyser

This section describes *AnIta*, a morphological analyser for Italian based on a large hand-written lexicon and two-level rule-based finite-state technologies. The motivations for such choice can be traced back, on the one hand, to the availability of a large electronic lexicon ready to be converted for such 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 models 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. [18, 20]).

The implementation of derivational phenomena in Italian considering the framework of two-level morphology has been extensively studied by [4]; 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 [14].

Considering the morphotactics combinations allowed for Italian, we have currently defined about 110,000 lemmas, 21,000 of which without inflection, 51 continuation classes to handle regular and irregular verb conjugations (following the proposal of [1] 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. We also added 3,461 proper nouns from person names, countries, cities and Italian politicians surnames to the AnIta lexicon.

Table 1. Some examples of AnIta analyses.

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

The availability of a large morphological analyser for Italian became very precious for developing a performant lemmatiser; the AnIta lexicon contains a very large quantity of Italian lemmas and is able to generate and recognise millions of wordforms and

assign them to a proper lemma (or lemmas). Testing the analyser coverage on CORIS, a large reference corpus of contemporary written Italian [17], 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.

Unfortunately, the morphological analyser cannot disambiguate the cases in which the wordform is ambiguous both from an orthographic and grammatical point of view (see [19] 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 [8] and applied to the De Mauro/Paravia online dictionary [9]. 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 table 2.

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	

2 Results and Discussion

Table 3 shows the lemmatisation task results: 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 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.

Table 3. EVALITA 2011 Lemmatisation Task results.

System	Lemmatisation Accuracy
1st Participant	99.06%
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%

References

1. Battista M., Pirrelli V.: Monotonic Paradigmatic Schemata in Italian Verb Inflexion. In: Proc. of COLING96, Copenhagen, pp. 77-82 (1996)
2. Battista M., Pirrelli V.: Una piattaforma di morfologia computazionale per l'analisi e la generazione delle parole italiane, ILC-CNR (2000)
3. Beesley, K.R., Karttunen, L. Finite State Morphology, CSLI Publications (2003)
4. Carota, F.: Derivational Morphology of Italian: Principles for Formalisation. Literary and Linguistic Computing, 21, pp. 41-53 (2006)
5. Cöltekin, C.: A Freely Available Morphological Analyzer for Turkish. In: Proc. of the 7th International Conference on Language Resources and Evaluation (LREC2010), Valletta, Malta (2010)
6. 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)
7. Delmonte, R.: Computational Linguistic Text Processing - Lexicon, Grammar, Parsing and Anaphora Resolution, Nova Science Publisher, New York (2009)
8. De Mauro, T.: Guida all'uso delle parole, Editori Riuniti, Roma (1980)
9. De Mauro, T.: Il dizionario della lingua italiana, Paravia (2000)
10. Gridach, M., Chenfour, N.: XMODEL: An XML-based Morphological Analyzer for Arabic Language. International Journal of Computational Linguistics, 1(2), pp. 12-26 (2010)
11. Hammarström, H., Borin, L.: Unsupervised Learning of Morphology. Computational Linguistics, 37(2), pp. 309-350 (2011)
12. Kiraz, G.A.: Computational Nonlinear Morphology: with emphasis on Semitic Languages, Cambridge University Press (2004)

13. Koskenniemi, K.: Two-level morphology: A general computational model for word-form recognition and generation. PhD Thesis, University of Helsinki (1983)
14. 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)
15. Pianta, E., Girardi, C., Zanoli, R.: The TextPro tool suite. In: Proc. of the 6th Language Resources and Evaluation Conference (LREC 2008), Marrakech (2008)
16. Roark, B., Sproat, R.: Computational Approaches to Morphology and Syntax, Oxford University Press (2006)
17. 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. and McEnery, T. (eds.), A Rainbow of Corpora: Corpus Linguistics and the Languages of the World, Lincom-Europa, Munich, 27-38 (2002)
18. 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)
19. Tamburini, F.: The EVALITA 2011 Lemmatisation Task, In: Working Notes of EVALITA 2011, 24th-25th January 2012, Rome, Italy (2012)
20. Tzoukermann, E., Libermann, M.Y.: A finite-state morphological processor for Spanish. In: Proc. of COLING'90, pp. 277-281 (1990)
21. Zanchetta, E., Baroni, M.: Morph-it! A free corpus-based morphological resource for the Italian language. In: Proc. Corpus Linguistics 2005, Birmingham (2005)