Italian Lemmatization by Rules with Getaruns

Rodolfo Delmonte

Department of Language Sciences
Ca' Bembo 1075 - 30123 Venezia
delmont@unive.it,
project.cgm.unive.it

Abstract. We present an approach to lemmatization based on exhaustive morphological analysis and use of external knowledge sources to help disambiguation which is the most relevant issue to cope with. Our system GETARUNS was not concerned with lemmatization directly and used morphological analysis only as backoff solution in case the word was not retrieved in the wordform dictionaries available. We found out that both the rules and the root dictionary needed amending. This was started during development and before testset was distributed, but not completed for lack of time. Thus the task final results only depict an incomplete system, which has now eventually come to a complete version with rather different outcome. We moved from 98.42 to 99.82 in the testset and from 99.82 to 99.91 in the devset. As said above, this is produced by rules and is not subject to statistical evaluation which may change according to different training sets. In this version of the paper we perform additional experiments with WordForm dictionaries of Italian freely available online.

Keywords: rule-based lemmatization, morphological analysis, semantically organized root-dictionary, semantic disambiguation.

1 Introduction

We present an approach to lemmatization[1] based on exhaustive morphological analysis and use of external knowledge sources to help disambiguation which is the most relevant issue to cope with. Our system GETARUNS [1,2,3] was not concerned with lemmatization directly and used morphological analysis only as fallback solution in case the word was not retrieved in the wordform dictionaries available. Lemmata were associated directly to wordforms and no provision was available for disambiguation. In fact, the shallow version of the system was only concerned with tagging for syntactic analysis. The deep system, on the contrary, is used only under the closed domain hypothesis and all information needed is generated, manually checked and used as is to produce semantic analysis. Thus, we have been obliged to work on a new complete version of the morphological analyser in order to generate

[1] This work has been partially funded by the PARLI Project (Portale per l'Accesso alle Risorse Linguistiche per l'Italiano – MIUR – PRIN 2008).

B. Magnini et al. (Eds.): EVALITA 2012, LNCS 7689, pp. 239–248, 2012.

best disambiguated lemmatized wordforms for the task, starting from what we had available.

We assume that the task of lemmatization in a morphologically rich language like Italian requires a rule-based approach to cope with the richness of wordforms produced which override by far 2million wordforms only for verb category. Generating appropriate wordform analysis and lemmata requires a complete morpheme list and a root dictionary adequately classified. Linguistic rules take both morphemes and roots classifications as input and implement a set of constraints to allow for recognition/generation of only legal wordforms and disallow illegal ones. Legal wordforms are typically a lot more than those actually present in Italian texts.

Lexical analysis in the GETARUNS system has been described extensively in a number of different papers presented in conferences in the past starting from the '80s (see [4,5,6,7]). Here we will concentrate on the system description rather than on the dictionaries and other resources used in the task. These will be briefly commented on in this section.

The system GETARUNS for Italian Lemmatization is composed of the following modules:

- a Root Dictionary made up of some 65,000 entries;
- a Dictionary of Invariable Wordforms including exceptional words like compounds with internal morphological variations, made up of 20,000 entries;
- a list of morphemes, which include 250 suffixes, 650 prefixes, 1050 derivational suffixes;
- precompiled lemmatized tagged wordforms included in separate lists, some of them with frequencies of occurrence for 75,000 entries – 28,000 of which with frequency of occurrence, these latter are derived from our Treebank called VIT (Venice Italian Treebank);
- a list of Italian wordforms with frequency of occurrence of 100,000 entries.

The algorithm for the Lemmatization Task is organized as follows:

1. Punctuation and other invariable words associated to categories which are not part of the evaluation are skipped in a first call;
2. Second and third call select the preceding one or two word context for the current word to analyse. The reason for introducing context of preceding words is dictated by the need to use redundant morphological information in determiners and modifiers preceding Nouns in order to help the disambiguation module;
3. Fourth call is the main call where the word is analysed and lemmatized. This will be explained in detail in a section below. However, this is only for words that are recognized at morphemic level for having at least a legal root and a legal suffix;
4. Fifth call is for words not recognized but still available in one of the wordform-lemmata list available;
5. Eventually, the guesser is activated, for those words that are not legally recognized: in this call, adjectives, verbs and nouns are analyzed according to their ending, disregarding the possible root. Depending on the suffix, specific rules are formulated to produce the adequate lemma in relation with lexical category.

2 The Algorithm for Morphological Analysis

In this section we will describe the algorithm for morphological analysis and the disambiguator phase. The algorithm is organized in the steps discussed below.

In the first step words ending in consonant are analysed and lemmatized directly. These words are not subject to disambiguation and the analysis ends up with just one possible interpretation. The information needed to process these words is either contained in a specialised list in the dictionary of invariable wordforms, or else they are recomposed with the missing apocoped vowel and then analysed directly. This applies to all types of functional words like demonstratives, possessives, indefinite adjectives and other similar categories. Also auxiliary and modal verbs are analysed in this part of the algorithm: their lemma is derived directly and is associated to each elided wordform. The same applies to nouns in case the word is included in the list of invariable wordforms where we see words referred to titles like "cavalier, dottor, ecc.", but also to words obeying general constraints for apocope in Italian. These rules are as follows:

- apocoped wordform must end with a sonorant consonant including "l, r, n", rarely "m"

Other wordform endings like "s" indicate that the word is not Italian and needs a different dictionary lookup. For this purpose we make available the two main dictionary for English and French that we organized for GETARUNS.

As for lexical verbs, their list is unpredictable and open: the complete wordform is passed to the main algorithm which however is called only once and is forced to produce the intended lemma as constrained by category.

In this step, all compound words are analysed in case they belong to a list of exception and can undergo unpredictable changes. This list includes all word composed with UOMO as second component, CAPO as first component and other similar cases. Special cases of plural are also included, those with "i" and double "e", for instance.

Second step is the main morphological algorithm which covers all other cases of wordforms, which in particular are not ending with consonant. Here, words are split into morphemes, notably root/theme and inflectional suffix, by stripping one character at a time starting from the right end of the word – i.e. reverting the order of the characters making up the wordform. The splitting process is made of two steps: at first characters are stripped and then reassembled into two components, then each component is checked for presence in the list of inflectional morphemes and roots. In case of success the process is interrupted and constrains are checked. An output analysis is then recovered if the splitting is legal, or rejected if the splitting is illegal. Splitting is then restarted from where it was interrupted by means of backtracking – which is freely made available in Prolog, our programming language.

Splitting continues up to a maximum suffix morpheme length of ten characters. All possible analysis are collected and then the output is passed to the disambiguation phase which will be described below in a separated section. Important subcases of this splitting process are constituted by verbal wordforms containing enclitics. Whenever such a case is spotted, the system enters a subroutine where the remaining part of the word is analysed and checked for consistency with the constraints. Other important

subcases are all wordforms belonging to irregular verbs. These are analysed by means of THEMEs and PREFIXEs and may have irregular endings too.

Third step regards all wordforms which have been rejected by the previous passage. The algorithm tries at first to split prefixes and then passes the remaining part of the word to the main algorithm. This is done recursively in order to collect all possible wordforms. At this point of the analysis also compound words with internal inflection are analysed and the corresponding lemma is recovered from the dictionary of invariable wordforms.

If this algorithm fails, the analysis continues by trying at first the opposite strategy: i.e. stripping all possible derivational suffixes which in turn may contain inflectional morphemes. This is done in three separate modalities: at first only derivational suffixes are searched and the remaining part of the word is searched in the root dictionary. Then, both prefixes and suffixes are searched and the remaining internal part of the word is searched as a root. Eventually only derivational suffixes are searched and the word type is guessed on the basis of the associated tag. However, basically verbs are not allowed to enter this part of the algorithm.

3 The Root Dictionary

The root dictionary is the heart of the morphological analyser. It is organised in twenty main lexical classes, as follows,

1. AGG adject.
2. AGGPP adject. participle past
3. AGGPR adject. participle present
4. AN adject.+noun attributive/predicative
5. ART article
6. AVV adverbial
7. CONG conjunction
8. COSU conjunction subordinate
9. CONGF conjunction coordinative sentential
10. EL element
11. INTER interjection
12. LOC locution (adverbial, conjunction, preposition)
13. N noun
14. NA noun+adject. predicative
15. NAPR noun+adject. participle present
16. PRE prefix
17. PREP preposition
18. PRON pronoun
19. SUFF suffix
20. V verb

then each class a certain number of subclasses which include information from all levels of computation. We will indicate below the number of morphological, syntactic and semantic subclasses but only one example per class, because of the lack of space -

but see Delmonte, Pianta (1996;1998) and Delmonte (1989) for irregular verb encoding rules.

1. AGG adjectival: 29 morphosyntactic subclasses
agg:co adj class -co antico

2. AGGPP participle past adjectival : one morphosyntactic subclass
aggpp:o classe -o moderato

3. AGGPR participle present adjectal: one morphosyntactic subclass
aggpr:e class -e mortificante

4. AN adject.+noun attributive/predicative: 14 morphosyntactic subclasses
an:comp adj+noun major maggiore

5. ART article: 2 morphological subclasses
art:def article def il

6. AVV adverbial (modifier of verb meaning): 12 morphosyntactic subclasses
avv:l adverbial locative qua

7.CONG conjunction (coordinates two phrases or sentences): 19 morphosyntactic subclasses
cong:av conj adversative bensì

8. CONGF conjunction sentential: 14 morphosyntactic subclasses
congf:av conj adversative viceversa

9. COSU conjunction subordinate: 6 morphosyntactic subclasses
cosu:av conj subord adversative anziché

10. EL element: 2 morphosyntactic subclasses
el:1 element first cloro

11. INTER interjection (can be used to build ellipsis): 1 morphosyntactic subclass
inter interjection diamine

12. LOC locution: 17 morphosyntactic subclasses
LOC AVV locution adverbial
loc:avv locution adverbial inintermediari

13. N noun: 46 morphosyntactic subclasses
n:a2:f noun fem class -a2 ala

14. NA noun+adject. predicative: 24 morphosyntactic subclasses
na:a:f noun+adj fem class -a femmina

15. NAPR noun+adject. participle present: 3 morphosyntactic subclasses

napr:e:f	noun+adj fem class -e	stimolante

16. PRE prefix: 4 morphosyntactic subclasses

pre	prefix	ri

17. PREP preposition: 2 morphosyntactic subclasses

prep	preposition	di

18. PRON pronoun : 24 morphosyntactic subclasses

pron:an	pron anaphoric	stesso

19. SUFF suffix: 13 morphosyntactic subclasses

suff:a	suffix adj	oica/o/che/ci

20. V verb: 53 morphosyntactic subclasses

v:1:cop	verb copulative 1.	sembrare

Overall there are 287 morphosyntactic subclasses which, as said above, also encode some semantics. Surely, they are used mainly to encode restrictions on root and word formation rules.

4 Lemmata Disambiguation

After lemmata have been associated to the wordform and category is matched with the entry tag, the disambiguation phase may start. This is obviously required only in case more than one different lemma is produced by the analysis. We need to distinguish cases related to nouns from other categories which require a different strategy. In particular, ambiguous verbforms are disambiguated on the basis of word frequency in large corpora: the two lemmata are compared on the basis of their frequency of occurrence and the most frequent is chosen. This is done simply on the basis of the fact that infrequent lemmata may correspond to archaic word meanings or simply orthography which are no longer used. As for adjectives, only masculine is allowed as lemma: in turn this may depend strictly on the class the adjective belongs to. Here we are referring to differences related to the inflectional suffix “i” interpreted as plural which may fit both into an “E” or “O” singular masculine ending. Information is collected in the root dictionary or else is derived from the Guesser.

Different lemmata may be generated at least in two cases:

- the wordform is a feminine gender word and has the same meaning of the masculine
- the wordform is a feminine gender word and does NOT have the same meaning of the masculine

In order to differentiate these two cases, roots in our dictionary have been separated. Thus the same ROOT may appear as separate entry twice or even three times in case of the existence of three different nominal endings. This has caused a careful search in the over 2000 entries that exhibited the problem, i.e. were classified as belonging to more than one nominal class. The problem was that in the majority of the cases, the referred meaning was not easily understandable because it belonged to some uncouth semantic domain and was as such not available in the high frequency dictionary of a normal Italian speaker. A search into online dictionaries was then required and being not always successful repeated.

Whenever the wordform was found semantically ambiguous on the basis of the meaning, the context was used as first disambiguator. In case a local determiner or modifier was encountered with a given gender, this was imposed on the following noun. Problems remained only for words which did not have any preceding disambiguating determiner. With these words we searched the wordform associated to the lemma in the frequency dictionary and decided to assign the most frequent lemma to the wordform.

However, this strategy did not always offer a satisfactory solution. One case is constituted by nouns referring to scientific branches of knowledge, as for instance "MATEMATICA, LOGICA, ARITMETICA, etc." when used in the feminine gender the choice was to keep that form also for the lemma, inspite of the possibility that the meaning would also refer to a person having the property of being such, which required the lemma in the masculine form.

5 Evaluation and Discussion

As said in the Abstract, when we submitted the results for the testset the work in the root dictionary had just started. Also some of the rules were missing, or were just incomplete. Work has continued slowly since then and the final results are much higher:

- TESTSET: from 98.42 we went up to 99.82
- DEVSET: from 99.82 we went up to 99.91

In one case we discovered that there was no rule in the algorithm to account for the plural form adjectives like LISCE/"smooth", MOSCE/"floppy" etc. and nouns like COSCE/thighs. In fact these words behave differently from other similar classes with a root ending with a palatal consonant because they require the addition of an "I" in the theme of the word. The root associate to these words must thus be "LISC", "MOSC" for the adjectives and "COSC" for the noun. Then a specific rule must associate an I to the theme in order to produce the singular form LISCIA/LISCIO, MOSCIA/MOSCIO, COSCIA.

However, mistakes are in many cases unavoidable because of the ambiguity present in the wordform and the difficulty in finding appropriate means to overcome it. Here below we present some classes of words which constitute impossible cases for disambiguation according to our approach, obviously.

CLASS 1.
Word Ending in E: 1st meaning Plural in E/ 2nd meaning Singular in E
POLTRONE (plural for POLTRONA/"armchair") – singular meaning "lazy person", VITE (plural for VITA/"life") – singular meaning "vine", PENE (plural for PENA/"pain") – singular meaning "cock", TESTE (plural for TESTA/"head") – singular meaning "witness", etc.
CLASS 2.
Word Ending Plural in HI: 1st meaning Singular in HIO/ 2nd meaning Singular in O
MARCHI plural for MARCHIO/"trade mark" – plural for MARCO/German currency Marc
CLASS 3.
Word Ending Plural in RI: 1st meaning Singular in IO/ 2nd meaning Singular in E
MARTIRI plural for MARTIRE/"martyr" – plural for MARTIRIO/"martyrdom", OSSERVATORI plural for OSSERVATORE/"observator" – plural for OSSERVATORIO/"observatory", ecc.
CLASS 4.
Word Ending Plural in NI: 1st meaning Singular in IO/ 2nd meaning Singular in E
QUARANTENNI plural for QUARANTENNE/"40-year-old-man" – plural for QUARANTENNIO/"40-year-period"
CLASS 5.
Word Ending Plural in INA: 1st meaning Singular in O/ 2nd meaning Singular in INA
TRENTINA meaning both a feminine inhabitant of Trento province (as such requiring a masculine lemma in O) and "a lot of thirty"
CLASS 6.
Word Ending Plural in INE: 1st meaning Singular in A/ 2nd meaning Singular in E
TENDINE meaning both an alteration of TENDA/"small curtains" and "(achille's) tendon"
CLASSE 7.
Word Ending Plural in I: 1st meaning Singular in O/ 2nd meaning Singular in E
FINI plural of FINE/"end" – plural of FINO/"fine", TESTI plural for TESTO/"text" – plural of TESTE/"witness", etc.

In order to check the quality of the baseline with the help of auxiliary resources, and also to test the hypothesis that claims the uselessness of morphological decomposition for the task of lemmatization - we decided to start updating and adapting two different main resources of Italian, which are available online freely - or perhaps in some cases, were available for download sometime ago. They are the following ones:

- MORPHIT WORDFORM Dictionary - University of Bologna
- Pisa University CoLFIS WORDFORM Dictionary

The resources have been manually checked for consistency and adapted to Prolog format. Then a wrapper for each category set has been produced in order to allow our system to use it conveniently. We will comment each resource and then discuss the results on both the Dev and the Test Set of the Evalita Lemmatization Task.

Morphit has 496,957 fully encoded and lemmatized entries which - once ported under Prolog - look like this,

```
mf(vacante,vacare,'VER','part+pres+s+f').
mf(vacantissima,vacante,'ADJ','sup+f+s').
mf(utopisti,utopista,'NOUN','m:p').
mf(te,te,'PROPERS','2+f+s').
mf(quanto,quanto,'PRO','WH-M-S').
mf(stessi,stare,'ASP','sub+impf+1+s').
mf(sti,questo,'DET','DEMO:m+p').
mff(veh,veh,'INT').
mff(velatamente,velatamente,'ADV').
mff(via,via,'PRE').
```

As can be seen, the morphology is accompanied by semantic information. We did some re-encoding in order to normalize some of the subclasses. We also separated invariable words - which have been given arity 3 - from variable words which have another slot for features.

In Table 1. below accuracy scores for the Dev and Test set are reported. As can be noticed, we did three runs each: first run with no additional information apart from using the same word as lemma in case it is missing from the entries of Morphit. We counted the missing words and they were 2656 in the testset, and 325 in the devset.

Table 1. Further evaluation carried on WordForm Dictionaries separated into Levels of information

	Testset	**Devset**
Level0	94.48	94.74
Level1	96.60	95.25
Level2	97.35	95.47
Level3	97.76	97.78

Levels refer to different amount of information made available to the lemmatizer: Level0 refers to the use of Morphit for all required categories, Noun-Verb-Adjective, and missing words are not lemmatized. Level1 allows for use of the same wordform as the corresponding lemma in case the word is not included in the dictionary. Level2 and Level3 make use of additional information coming either from our dictionaries, or eventually from morphological guessing. We did not include the morphological analyser for obvious reasons, but the guesser that we also use for out of vocabulary words.

Pisa CoLFIS WordForm Dictionary, has 182,357 entries encoded with frequency of occurence - dispersion and other interesting frequency related data - for both wordform and lemma, where each entry looks like this,

pli(cuccioli, 'S', cucciolo, 'S').
pli(cucco, 'V', cuccare, 'V').
pli(cui, 'N', cui, 'N').

Given the reduced number of wordform we expect a lower performance at level 0, which is what we found:

DevSet Accuracy: 93.54
TestSet Accuracy: 93.44

So eventually it is important to consider the amount of information already available when building a lemmatizer, so that no unneeded extra work is done by the morphological analyzer. However, the information encoded needs to be carefully checked not to induce mistakes in the lemmatization process.

To conclude, we assume that precompiled resources may be useful as long as they are well organized and manually checked.

References

1. Delmonte, R., Mian, G.A., Tisato, G.: Un riconoscitore morfologico a transizioni aumentate, pp. 100–107. Atti Convegno Annuale A.I.C.A, Firenze (1985)
2. Delmonte, R.: Computational Morphology for Italian. In: Delmonte, R., Ferrari, G., Prodanof, I. (eds.), Studi di Linguistica Computazionale, vol. I, pp. 109–162. Unipress, Padova (1988)
3. Delmonte, R.: Verbi irregolari: una analisi computazionale. In: Delmonte, R. (ed.) Lessico, Strutture e Interpretazione - Studi Linguistici Applicati I, ch. I, pp. 3–59. Unipress, Padova (1989)
4. Delmonte, R.: Lexical Representations: Syntax-Semantics interface and World Knowledge. In: Rivista dell'AI*IA, pp. 11–16. Associazione Italiana di Intelligenza Artificiale, Roma (1995)
5. Delmonte, R., Pianta, E.: IMMORTALE - Analizzatore Morfologico, Tagger e Lemmatizzatore per l'Italiano. In: Atti Convegno Nazionale AI*IA Cibernetica e Machine Learning, Napoli, pp. 19–22 (1996)
6. Delmonte, R.: Rappresentazioni lessicali e linguistica computazionale. In: Atti, S.L.I. (ed.) Lessico e Grammatica - Teorie Linguistiche e applicazioni lessicografiche, Roma, Bulzoni, pp. 431–462 (1997)
7. Delmonte, R., Pianta, E.: Immortal: How to Detect Misspelled from Unknown Words. In: BULAG, PCUF, Besançon, pp. 193–218 (1998)
8. http://linguistica.sns.it/CoLFIS/Home.html
9. http://dev.sslmit.unibo.it/linguistics/morph-it.php