10.12871/CLICIT2014141

Part-of-Speech tagging strategy for MIDIA: a diachronic corpus of the Italian language

Claudio Iacobini
University of Salerno
ciacobini@unisa.it

Aurelio De Rosa
University of Salerno
aurelioderosa@gmail.com

Giovanna Schirato
University of Salerno
giovanna.schirato@gmail.com

Abstract

English. The realization of MIDIA (a balanced diachronic corpus of written Italian texts ranging from the XIII to the first half of the XX c.) has raised the issue of developing a strategy for PoS tagging able to properly analyze texts from different textual genres belonging to a broad span of the history of the Italian language. The paper briefly describes the MIDIA corpus; it focuses on the improvements to the contemporary Italian parameter file of the PoS tagging program Tree Tagger, made to adapt the software to the analysis of a textual basis characterized by strong morpho-syntactic and lexical variation; and, finally, it outlines the reasons and the advantages of the strategies adopted.

Italiano. *La realizzazione di MIDIA (un corpus diacronico bilanciato di testi scritti dell'italiano dal XIII alla prima metà del XX secolo) ha posto il problema di elaborare una strategia di PoS tagging capace di analizzare adeguatamente testi appartenenti a diversi generi testuali e che si estendono lungo un ampio arco temporale della storia della lingua italiana. Il paper, dopo una breve descrizione del corpus MIDIA, si focalizza sui cambiamenti apportati al file dei parametri dell'italiano contemporaneo per il programma di PoS tagging Tree-Tagger al fine di renderlo adeguato all'analisi di una base testuale caratterizzata da una forte variazione morfosintattica e lessicale, e evidenzia le motivazioni e i vantaggi delle strategie adottate.*

1 Introduction

The realization of MIDIA, a balanced diachronic corpus of Italian, raised the issue of the elaboration of a strategy of analysis of texts from different genres and time periods in the history of Italian. This temporal and textual diversity involves both a marked graphic, morphological and lexical variation in word forms, and differences in the ordering of the PoS. The program chosen for the PoS tagging is Tree Tagger (cf. Schmid 1994, 1995), and the parameter file, made of a lexicon and a training corpus, is the one developed by Baroni et al (2004) for contemporary Italian. The strategy we developed for the adjustement of the PoS tagging to different diachronic varieties has been to greatly increase the lexicon with a large amount of word forms belonging predominantly to Old Italian, and not to retrain the program with texts belonging to previous temporal stages. This solution turned out to be economical and effective: it has allowed a significant improvement of the correct assignment of PoS for texts both old and modern, with a success rate equal to or greater than 95% for the tested texts, and an optimal use of human resources.

2 MIDIA: a brief description

MIDIA (an acronym for Morfologia Italiana in Diacronia) is a balanced diachronic corpus of written Italian texts, fully annotated with the indication of the lemma and the part of speech. The corpus goes from the beginning of the thirteenth to the first half of the twentieth century.

Periodization is based on important linguistic, literary and cultural facts of Italian history. Five time periods have been distinguished: 1) 1200-1375 formation of Tuscan-centered Old Italian; 2) 1376-1532 affirmation of Italian outside Tuscany; 3) 1533-1691 standardization of Italian in the late Renaissance, Mannerist and Baroque periods; 4) 1692-1840 the birth of modern Italian: the age of Arcadia, the Enlightenment and Romanticism; 5) 1841-1947 the language of Italian political unification.

Texts belonging to seven genres have been collected: expositive prose; literary prose; normative and juridical prose; personal prose; scientific prose; poetry; spoken language mimesis. For each time period and genre 25 texts were selected. A section of 8000 tokens was extracted from each text; for a total of more than 7.5 million tokens.

The search tool we built, in the form of a web application, allows an easy extraction of the data, particularly devised for the study of word-formation in Italian from a diachronic point of view, but also usable for several other types of linguistic investigation. MIDIA can be queried for forms or lexemes also through the use of regular expressions, the search can be refined through the identification of word forms, lexemes or PoS that precede or follow the queried string, and through the use of metadata concerning time period, genre, author, and work.

Different types of outcome can be obtained. The default result shows the selected string in context (the value of 10 left and 10 right forms can be increased or decreased) together with the indication of PoS, lexeme, the metadata concerning author and work, and the ID of the file containing about 8,000 token texts from which the selected item is taken. Other outcomes consist of: distribution tables indicating the number of occurrences of the selected item distinguishing genres and periods; frequency lists showing the number of occurrences of the selected item divided in form, PoS and lemma; graphs and charts showing time evolution of the selected item according to author, genre, and period. All the types of outcome can be viewed online and downloaded in CSV format.

MIDIA is the outcome of the Prin project "The history of word-formation in Italian" funded by the Italian Ministry of Education University and Research. The corpus is freely available at the URL http://www.corpusmidia.unito.it/.

3 PoS tagging strategy for a diachronic corpus of the Italian language

The software we used to associate a part of speech to each word form of our corpus is Treetagger (cf. Schmid 1994, 1995). The application of the Tree Tagger software to a language involves the identification of a Tagset, the creation of a lexicon containing the a priori tag probabilities for each word, and a Tagged Corpus representing the (variety of the) language that is to be analyzed.

We started the automatic annotation with part-of-speech tags using the source files underlying the parameter file for contemporary Italian made by Baroni et al. (2004), which consists of a training corpus of about 115,000 tokens taken from the newspaper La Repubblica (years 1985-2000), and a lexicon which amounts to approximately 220,000 tokens (we thank Marco Baroni for his contribution to the realization of our project).

Our case presents special problems because of the variety of genres and the time span of the texts of the corpus (about PoS tagging of diachronic corpora, cf. Dipper et al. 2004, Martineau 2008, Sánchez-Marco et al. 2010, Stein 2008). We began to test the contemporary Italian TreeTagger (ContIt TT) on two literary prose texts of the first period (1200-1375) of our corpus (taken from Dante's *Vita Nuova* and Dino Compagni, *Cronica delle cose occorrenti ne' tempi suoi*) in order to figure out the problems that the program had with Old Italian texts. The results have been manually checked in order to find recurring mistakes and to think about some possible solutions for the improvement of PoS tagging.

The result of POS tagging on the two texts of the first period was then compared with that of a literary prose text of the most recent period (1841-1947) of our corpus: Italo Svevo, *La coscienza di Zeno*. As expected, the error rate of ContIt TT, fully satisfactory for modern texts (about 5%), was higher for Old Italian literary prose (about 13%). In addition, error analysis reveals that wrong assignments mainly concern PoS (exp. adjectives and verbs) of particular interest for the study of word-formation, for which the MIDIA corpus is especially conceived.

As is known, TreeTagger is a probabilistic PoS tagger that gives to each token of a text PoS and lemma information. The assignment of a particular PoS to each word form depends on the matching with a form present in the lexicon associated with the probabilities of co-occurrence of a PoS with other adjacent according to the information about PoS sequences obtained from the training corpus.

The strategy we adopted to cope with our diachronic corpus was to strongly enrich the contemporary Italian lexicon (that is, the list of forms with specification of PoS and lemma) and not to train it on a widened corpus to which were added Old Italian texts (cf. Gaeta et al. 2013). Our expectation was that PoS tagging of the diachronic corpus could be significantly improved even without adding to the training corpus examples of the typical syntactic patterns found in Old Italian texts.

The reason behind this decision is twofold. On the one hand we took a theoretical and methodological stance: we were confident that by adding more forms (especially those more typical of older texts) we could significantly improve the results of the analysis, i.e. to have a better "syntactic" analysis through more detailed word recognition. On the other hand we took a cautious position: since ContIt TT already had fairly good results also with Old Italian texts, we have preferred avoiding to alter the distribution of the sequence of PoS on which the program was set (by adding a training corpus made of early texts), especially considering that MIDIA corpus is made not only of texts belonging to Old Italian, but to the entire time span of the history of Italian.

Our expectation was that the recognition of word forms would significantly help the recognition of sentences, i.e. the recognition of sequences of PoS elements, and this was what happened (as we will show in section 4).

The enrichment of MIDIA Tree Tagger (MIDIA TT) lexicon results from the addition of about 230,000 word forms mainly dating from the XIV to the XVI c. (MIDIA TT lexicon actually counts about 550,000 forms).

For the implementation of the lexicon, in a first step we have made use of the available philological resources: word lists, lists of names, critical editions, glossaries and digital corpora (Corpus Taurinense TLIO); later, comparing the lexicon increased in this way with the set of forms used in the texts of the MIDIA corpus, we selected those absent from the lexicon, favoring forms with higher frequency and morphological variance, and we tagged them with a semiautomatic procedure according to the format required by Tree Tagger, paying particular attention to the homographs that would have troubled the recognition mechanisms of the program (for example, proper names were not included that would have generated ambiguity overlap with common names: *Prato, Potenza, Monaco, Fiume, Riga, Spine, Spira, Angelo, Norma, Nunzio, Leone,* etc.; with verbs: *Segna, Segni, Giura, Vendi*; or with numerals: *Cento*). For the same reason we have reduced the Tagset analyticity by suppressing the distinction between adjectives and pronouns for demonstratives, indefinites, numerals, possessives, interrogatives.

4 Checking the results of MIDIA PoS tagging and error analysis

In order to evaluate the performance of MIDIA TT, we have selected one text of literary prose for each of the time periods of the corpus, and for each text we prepared a gold standard PoS assignment through a thorough manual review revised and discussed within our research group.

These gold standards form the benchmark for the performance evaluation of the ContIt TT and MIDIA TT programs (the number of tokens manually checked for PoS assignment is 52,952).

Table 1: See appendix

Table 1 compares the number and the percentage of errors in ContIt TT and MIDIA TT PoS tagging for literary texts belonging to the five time periods. As may be noted, MIDIA TT has significantly better results than those of ContIT TT especially in the first periods; furthermore, we can notice that the result of MIDIA TT in period 1 is better than that of ContIt TT in period 5.

Tables 2 and 3 show some of the typical errors of ContIt TT (highlighted in bold) compared with MIDIA TT correct PoS tagging in texts belonging to the first period.

Table 2: See appendix

Table 3: See appendix

ContIt TT PoS tagging errors reported in bold in Table 2 are very likely to be attributed to the recognition of *ser* (antiquated form for 'mister', but similar in form to the verb *essere* 'to be') as a Noun, which results in the assignment of the form *dove* to the PoS WH instead of to Conjunction; the absence in ContIt lexicon of *giacea* and the proximity of this form to a proper noun (*Ciappelletto*) causes the erroneous tagging of this Verb to the adjectival class. Similarly, the form *allato* is recognized as a past participle (probably because of the final string), while *postoglisi* is not recognized as a past participle because of the combination of enclitic forms. MIDIA lexicon contains all these verb forms and allows the correct attribution of the PoS Conjunction to the word form *dove*, although in the lexicon this form corresponds to three different PoS (Noun, Adverb, and Conjunction).

In table 3 the ambiguity of *magnifico* (Noun, Adjective, and Verb) and the absence in ContIt lexicon of the word form *suggeritole* causes the error in the assignment of PoS of these forms and of the adjacent word *quadretto*. From this brief error analysis, we may conclude that the failure to recognize word forms triggers a cascade effect of PoS assignment on nearby words, whereas a rich lexicon increases the possibility of a correct PoS assignment also for words that are not listed in the lexicon.

Table 4 shows the PoS with a higher percentage of errors in the text of the first period used as gold standard for PoS assignment (the column GS shows the expected number of tokens for each PoS; the left column of both ContIt TT and Midia TT shows the difference from GS, the right column the percentage of errors for each PoS assignment).

Table 4: See appendix

The errors in MIDIA TT are concentrated in clitics, auxiliary and modal verbs (which generally are still recognized as verbs). The nouns do not present serious problems either in MIDIA TT or in ContIt TT, while the latter has a high error rate in the adjectives, verbs and adverbs; the difficulty in recognizing the members of these PoS is probably due to their high graphic and morphological variation not accounted in ContIt lexicon. The main errors in the PoS tagging of Old Italian in MIDIA TT can be traced in part to the decision not to train MIDIA TT with texts of this period. The main differences that distinguish modern and contemporary Italian from Old Italian concern primarily the syntactic structure; among the syntactic differences, one of the most notable is the possibility to interpose nominal arguments between modal and auxiliary verbs and the main verb, and a greater freedom of clitic position (Renzi and Salvi, 2010; Dardano, 2013). The criterion of adding word forms to the lexicon cannot cope with these difficulties, while it has proved to be adequate for many other variation factors, such as lexical and morphological differences, and also the different positions of the main verbs or of the nominal constituents. The overall positive result on the texts of all the periods made us decide to maintain our choice. Moreover, the enriched lexicon can still provide a useful starting point for those just interested in the texts of Old Italian, who want to train a Tree Tagger parameter file specialized for these texts.

Table 5: See appendix

Table 5 compares auxiliaries, clitics and verbs PoS tagging in period 1 and 5. It shows that verb recognition is stable in the two periods for MIDIA TT, while the correct assignment of clitics and auxiliaries strongly improves in the most recent period for both MIDIA TT and ConIT TT. The good results in verb recognition already performed by MIDIA TT in period 1 may be attributed to the strong enrichment of the lexicon (cf. the high percentage of errors of Cont It TT in period 1), the differences in auxiliaries and clitics can be explained with changes in the syntactic order in the two periods of the Italian language under examination.

5 Conclusions

The strategy we devised to develop MIDIA PoS tagging for the analysis of texts belonging to different time periods and textual genres than that for which it was originally trained has proved to be successful and economical. Human resources have been concentrated on enriching the lexicon and on the review of automatic lexeme and PoS assignment.

Our results show that a larger lexicon improves the analysis also for words adjacent to those recognized by the matching with the word forms listed in the lexicon. This has some interesting consequences both on the strategies for text tagging and on the implementation of the program Tree Tagger for the analysis of texts with a great range of variation.

We plan to further enrich MIDIA lexicon by adding word forms from the corpus not yet listed in the lexicon.

References

Baroni Marco, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi,. Guy Aston, and Marco Mazzoleni. 2004. Introducing the "la Repubblica" corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian. *Proceedings of LREC 2004*.

Dardano Maurizio (ed.). 2013. *Sintassi dell'italiano antico*. Carocci, Roma.

Dipper Stefanie, Faulstich Lukas, Leser Ulf and Lüdeling Anke. 2004. Challenges in Modelling a Richly Annotated Diachronic corpus of German. *Proceedings of the LREC Workshop on XML-based Richly Annotated Corpora*, Lisbon, Portugal: 21-29.

Gaeta Livio, Claudio Iacobini, Davide Ricca, Marco Angster, Aurelio De Rosa, and Giovanna Schirato. 2013. MIDIA: a balanced diachronic corpus of Italian. *Conference held at 21st International Conference on Historical Linguistics (Oslo, 5-9 August 2013)*.

Martineau France. 2008. Un corpus pour l'analyse de la variation et du changement linguistique. *Corpus* [En ligne] 7: 136-155. URL : http://corpus.revues.org/1508

Renzi Lorenzo and Giampaolo Salvi. 2010. Italiano antico. In:. *Enciclopedia dell'italiano*. Roma, Istituto dell'Enciclopedia Italiana: 713-716.

Schmid Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing, Manchester, UK*.

Sánchez-Marco Cristina, Boleda Gemma, Fontana Josep Maria, Domingo Judith. 2010. Annotation and representation of a diachronic corpus of Spanish. *Proceedings of the International Conference on Language Resources and Evaluation*, 17-23 May, Valletta, ELRA: 2713-2718.

Schmid Helmut. 1995. Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland*.

Stein Achim. 2008. Syntactic Annotation of Old French Text Corpora. *Corpus* [En ligne] 7: 157-171. URL : http://corpus.revues.org/1510

Appendix

Period	**ContIt TT**		**MIDIA TT**	
1	1260	13.24%	478	5.02%
2	1117	9.87%	507	4.48%
3	1904	15.37%	493	3.98%
4	902	8.62%	272	2.60%
5	568	6.13%	220	2.38%
average		**10.86%**		**3.72%**

Table 1: Number and percentage of PoS tagging errors in the five time periods of the corpus MIDIA.

TEXT	**ContIt TT**	**MIDIA TT**
nella	ARTPRE	ARTPRE
camera	NOUN	NOUN
dove	**WH**	CON
ser	**VER:fin**	NOUN
Ciappelletto	NPR	NPR
giacea	**ADJ**	VER:fin
e	CON	CON
allato	**VER:ppast**	ADV
postoglisi	**NOUN**	VER:ppast:cli
a	PRE	PRE
sedere	VER:infi	VER:infi

Table 2: Error analysis (first period texts).

TEXT	**ContIt TT**	**MIDIA TT**
dipinse	VER:fin	VER:fin
un	ART	ART
magnifico	**NOUN**	ADJ
quadretto	**VER:fin**	NOUN
suggeritole	**NOUN**	VER:ppast:cli
dalla	ARTPRE	ARTPRE
mia	DET:poss	DET:poss
malattia	NOUN	NOUN

Table 3: Error analysis (first period texts).

POS	**GS**	**ContIt TT**		**MIDIA TT**	
ADJ	381	**166**	**43.6 %**	23	6.0 %
ADV	652	**132**	**20.3 %**	5	0.8 %
AUX	187	**78**	**41.7 %**	**61**	**32.6** %
CLI	287	**57**	**19.9 %**	**73**	**25.4** %
CON	565	41	7.3 %	10	1.8 %
DET	905	61	6.7 %	38	4.2 %
NOUN	1402	49	3.50 %	45	3.2 %
PRE	1080	79	7.31 %	1	0.1 %
PRO	542	43	7.9 %	1	0.2 %
VER	**1432**	**342**	**33.9** %	81	5.6 %
VER2	**134**	**46**	**34.3** %	**51**	**38.1** %

Table 4: PoS tagging errors (first period).

PoS	**Period 1**				
	GS	**ContIT TT**		**MIDIA TT**	
AUX	187	78	41.7%	61	32.6%
CLI	287	57	19.9%	73	25.4%
VER	1432	486	33.9%	81	5.6%
	Period 5				
	GS	**ContIT TT**		**MIDIA TT**	
AUX	213	6	2.8%	9	4.2%
CLI	342	49	14.3%	27	7.9%
VER	1476	151	10.2%	69	4.7%

Table 5: PoS tagging errors for auxiliaries, clitic and verbs in period 1 and 5.