PoS Taggers in the Wild: A Case Study with Swiss Italian Student Essays

Daniele Puccinelli, Silvia Demartini, Aris Piatti, Sara Giulivi, Luca Cignetti, Simone Fornara
University of Applied Sciences and Arts of Southern Switzerland (SUPSI)
{daniele.puccinelli, silvia.demartini, aris.piatti,
{sara.giulivi, luca.cignetti, simone.fornara}@supsi.ch

Abstract

English. State-of-the-art Part-of-Speech taggers have been thoroughly evaluated on standard Italian. To understand how Part-of-Speech taggers that have been pre-trained on standard Italian fare with a wide array of language anomalies, we evaluate five Part-of-Speech taggers on a corpus of student essays written throughout the largest Italian-speaking area outside of Italy. Our preliminary results show that there is a significant gap between their performance on non-standard Italian and on standard Italian, and that the performance loss mainly comes from relatively subtle tagging errors within morphological categories as opposed to coarse errors across categories.

Italiano. *Gli strumenti di Part-of-Speech tagging più rappresentativi dello stato dell'arte sono stati analizzati a fondo con l'italiano standard. Per capire come strumenti pre-addestrati sull'italiano standard si comportano in presenza di un'ampia gamma di anomalie linguistiche, analizziamo le prestazioni di cinque strumenti su di un corpus di elaborati redatti da studenti della scuola dell'obbligo nella Svizzera Italiana. I nostri risultati preliminari mostrano che esiste un notevole divario tra le prestazioni sull'italiano non-standard e quelle sull'italiano standard, e che la perdita di prestazioni deriva principalmente da errori di tagging relativamente sottili all'interno delle categorie grammaticali.*

1 Introduction

The goal of this paper is to present the preliminary results of the evaluation of a set of state-of-the-art Part of Speech (PoS) taggers on the DFA-TIscrivo corpus of Italian-language (L1) K-12 student essays from schools in the Italian-speaking part of Switzerland. The DFA-TIscrivo corpus represents an example of non-standard Italian[1] because its contributors are young students with a poor command of the Italian language living in the largest Italian-speaking area outside of Italy, and therefore prone to regionalisms as well as orthographic mistakes.

The key research question at this stage is how well state-of-the-art PoS taggers that were pre-trained on standard Italian cope with a specific flavor of non-standard Italian. It would of course be possible to retrain all these tools on texts with similar properties as the ones in our corpus, but at this stage in our work this is not possible due to the overly small size of the available annotated data. In turn, using pre-trained models gives us a twofold advantage: it allows us to obtain a performance baseline on non-standard Italian, and it makes it possible to directly compare our performance metrics to previously published results (obtained with the same models we use). While our work is still in progress and the results reported herein are preliminary in nature, we can already share several notable observations.

2 Related Work and PoS taggers under test

There have been various recent efforts focused on social media within the scope of EVALITA 2016 (Bosco et al., 2016), whose goal was the domain adaptation of PoS-taggers to Twitter texts. Notable contributions include (Cimino and Dell'Orletta, 2016), whose authors propose a PoS tagging architecture optimized to process Italian-language tweets. While we do acknowledge the need for

[1] http://www.treccani.it/enciclopedia/italiano-standard_(Enciclopedia-dell%27Italiano)/

domain adaptation with non-standard texts, we ask a more basic question: if we perform no domain adaptation and simply deploy general-purpose PoS taggers *in the wild*, how do they fare? We use K-12 student essays as our flavor of non-standard Italian. Although such texts are beset with all sorts of anomalies, they can still be processed them with general purpose taggers, unlike far more unstructured and unconventional texts such as tweets. While similar studies have been conducted for other languages, such as German (Giesbrecht and Evert, 2009), to the best of our knowledge this is the first study of the accuracy of general-purpose PoS taggers *in the wild* for the Italian language. Our selection of state-of-the-art general purpose PoS taggers is based on their popularity with the research community and the availability of ready-to-use software versions.

TreeTagger (1994). The popular TreeTagger (Schmid, 1994) tool uses decision trees to estimate transition probabilities based on context. Decision trees were extremely popular for PoS tagging in the 1990s, when more sophisticated machine learning tools such as neural networks were still too computationally demanding given the relatively limited resources available at the time. TreeTagger actively addresses the issues encountered by earlier probabilistic PoS taggers with rare words with a very low (but non-zero) probability of occurrence. The use of decision trees enables TreeTagger to account for context, whose nature is not restricted to $n-$grams, but also to allowed/disallowed tag sequences.

UD-Pipe (2014). UD-Pipe (Straka et al., 2016) is a language-agnostic natural language processing (NLP) pipeline developed within Universal Dependencies, whose focus is the development of a treebank annotation scheme that can work consistently across multiple languages. UD-Pipe's PoS tagger uses the Morphological Dictionary and Tagger MorphoDiTa (Straková et al., 2014), developed at Charles University in Prague, Czech Republic. MorphoDiTa uses the averaged perceptron PoS tagger described in (Spoustová et al., 2009) and based on (Collins, 2002).

Tint (2016). *The Italian NLP Tool* (Palmero Aprosio and Moretti, 2016) is an NLP pipeline for the Italian language based on Stanford CoreNLP (Manning et al., 2014). Tint's PoS tagger is based on the Stanford Log-linear Tagger (Toutanova et al., 2003), which leverages maximum entropy PoS tagging (Toutanova and Manning, 2000). Given a word and its context (other words in the sentence and their tags), maximum entropy PoS tagging assigns a probability to every tag in a predefined tagset, eventually enabling the estimation of the probability of a tag sequence given a word sequence. Out of all the possible distributions that satisfy a set of constraints, the one with maximum entropy is chosen, as it represents the most non-committal assignment of probabilities that meets the constraints (Ratnaparkhi, 1996).

Syntaxnet (2016). Various recent efforts focus on the application of recurrent neural networks to PoS tagging and dependency parsing (Ling et al., 2015), but it is shown in (Andor et al., 2016) that recurrence-free feed-forward networks can work at least as well as recurrent ones if they are globally normalized; this is the guiding principle behind PoS tagging in Syntaxnet (syn, 2016), a neural network NLP framework that is built on top of Google's popular TensorFlow machine learning framework (Abadi et al., 2016). Syntaxnet employs beam search, which serves to maintain multiple hypotheses, and global normalization with a conditional random field (CRF) objective, which avoids label bias issues (typically reported in locally normalized models). PoS tagging in Syntaxnet is heavily inspired by (Bohnet and Nivre, 2012) and relies on the close integration of PoS tagging and dependency parsing. A pre-trained English language model whimsically called *Parsey McParseface* was released along with Syntaxnet in May 2016 and a pre-trained model for the Italian language was released in August 2016 as one of *Parsey's Cousins*.

DRAGNN (2017). In March 2017 Google released a Syntaxnet upgrade based on Dynamic Recurrent Acyclic Graphical Neural Networks (DRAGNN) (Kong et al., 2017) along with the *Parseysaurus* set of pre-trained models (Alberti et al., 2017) that was developed for the CONLL 2017 shared task. PoS tagging in DRAGNN (Kong et al., 2017) is based on (Zhang and Weiss, 2016), which closely integrates PoS tagging and parsing in a novel fashion (specifically, the continuous hidden layer activations of the window-based tagger network are fed as input to the transition-based parser network). The tagger works token by token, extracting features from a window of tokens around the target token. It has a fairly standard

structure with embedding, hidden, and softmax layers.

3 The DFA-TIscrivo corpus

The DFA-TIscrivo corpus has been prepared within the projects TIscrivo (2011-2014) and TIscrivo 2.0 (2014-2017) projects[2], both funded by the Swiss National Science Foundation. The goal of the projects is to paint an accurate picture of the writing skills of primary school and lower secondary school in Southern Switzerland in order to describe the variety of language written at school and to propose new teaching practices to improve writing skills in compulsory education (Cignetti et al., 2016). Other studies with some similarities to the TIscrivo projects include projects focused on texts by L1 or L2 learners such as ISACCO (Brunato and Dell'Orletta, 2015), CItA (Barbagli et al., 2015)(Barbagli et al., 2016), and KoKo (Abel et al., 2016).

The DFA-TIscrivo corpus is a balanced corpus collected in 56 Italian-speaking primary and lower secondary schools from Southern Switzerland. It contains 1735 narrative-reflective essays (742 from primary, 993 from secondary school), transcribed but not normalized, and accompanied by sociolinguistic metadata (age, gender, school and class, linguistic information). It amounts to about 390,000 tokens. Lexical data were initially lemmatized and PoS tagged using TreeTagger (with the Italian parameters by Marco Baroni) and are being manually revised. Furthermore, we are manually annotating orthographic, morphological and lexical main types of error, multi-word expressions, peculiar lexicon of Italian only used in Southern Switzerland and foreign words. A key project goal is to build up a dictionary of the Italian language as it is written in Southern Switzerland (Cignetti and Demartini, 2016)(Fornara et al., 2016) as an online resource useful both to scholars and to teachers.

4 Methodology and Performance Analysis

We run the five taggers on the corpus and compare their output to a manually tagged ground truth. We note that, at the time of writing, the analysis is restricted to a subset of the DFA-TIscrivo corpus that has been manually PoS-tagged and is limited to essays written by fifth graders. We use the ISST-TANL-PoS reference tagset[3] based on Universal Dependencies.

	Accuracy
TreeTagger	0.84
UD-Pipe	0.79
Tint	0.83
Syntaxnet	0.83
DRAGNN	0.84

Table 1: Overall PoS tagging accuracy for each tool on the DFA-TIscrivo corpus.

We begin by assessing the tagging accuracy of the five PoS taggers under test on the DFA-TIscrivo corpus. We compute the tagging accuracy as the ratio of correctly tagged parts of speech with respect to the aforementioned manually tagged ground truth. While the ground truth isolates out multiword expressions, none of the tools are able to do that, so all multiword expressions are considered to be mistagged and every multiword expression counts as one single miss. Verbal enclitics are not considered and the corresponding verbs are expected to be tagged simply as verbs. Our results are shown in Table 1; we see that UD-Pipe trails behind and falls below the 0.8 mark, while the other four taggers under test offer a similar performance, with TreeTagger slightly ahead of the pack. All these taggers reportedly perform above the 95% mark on standard Italian.

Tables 2-6 contain the confusion matrices of the PoS taggers under test based on the ISST-TANL coarse-grained tags. Row i shows the ground truth for tag i and column k shows the frequency with which it is tagged as k. To abstract away from how individual taggers address prepositional article, we merge the tags for prepositions (E) and articles (R) into a super-tag ER. We also merge the tags for adjectives (A) and determiners (D) because determiners may be viewed as a category of adjectives in Italian. We only show the tags that occur most often (which is why some rows/columns do not add up to one). We note that TreeTagger outperforms all other taggers with AD while lagging behind all of them with P (pronouns) and C (conjunctions), often tagged as P or B (adverbs). TreeTaggers also performs remarkably well with verbs (V).

[2] http://dfa-blog.supsi.ch/DFA-TIscrivo~/la-ricerca/

[3] http://www.italianlp.it/docs/ISST-TANL-POStagset.pdf

	AD	B	C	ER	P	S	V
AD	0.95	0.03	0.01	0	0	0	0
B	0.02	0.88	0	0	0.02	0.04	0.04
C	0.01	0.07	0.76	0.01	0.15	0	0
ER	0.08	0	0	0.92	0	0	0
P	0.05	0	0.01	0	0.79	0.01	0
S	0.04	0	0	0	0	0.93	0.03
V	0	0	0	0	0	0.01	0.99

Table 2: Tree Tagger confusion matrix.

	AD	B	C	ER	P	S	V
AD	0.77	0.04	0	0	0	0.04	0.01
B	0.04	0.86	0	0.01	0	0.07	0.02
C	0	0.06	0.91	0.02	0	0.01	0
ER	0	0	0	1	0	0	0
P	0.01	0.01	0.02	0.02	0.93	0.01	0
S	0.02	0.01	0	0	0	0.96	0.01
V	0.02	0	0	0	0.01	0.03	0.94

Table 3: UD-Pipe confusion matrix.

	AD	B	C	ER	P	S	V
AD	0.75	0	0	0	0.15	0	0
B	0.06	0.88	0	0.01	0	0.03	0.02
C	0	0.09	0.89	0.01	0	0	0
ER	0	0	0	0.99	0	0	0
P	0.02	0	0	0.03	0.88	0.01	0
S	0.03	0	0	0	0	0.94	0.03
V	0.01	0	0	0	0	0.01	0.92

Table 4: Tint confusion matrix.

We have also studied the confusion matrices within the V category (not shown), noting that TreeTagger performs remarkably better than the others with respect to principal verbs (0.97 accuracy while the others are right around the 0.9 mark). and modal verbs (0.94 versus 0.81 for UD-Pipe and TINT and a disappointing 0.75 for both Syntaxnet and DRAGNN). All taggers perform equally poorly with auxiliary verbs (accuracy just above the 0.8 mark in all cases). Aside from Tint, which does not provide morphological information (at least in the version we used), all taggers do well with finite verbs (> 0.97, with UD-Pipe trailing behind at 0.95). While TreeTagger and UD-Pipe perform at the same level of accuracy for both finite and non-finite verbs, Syntaxnet and DRAGNN barely go beyond the 0.9 mark with the latter.

	AD	B	C	ER	P	S	V
AD	0.75	0.01	0	0	0.03	0.05	0.02
B	0.01	0.88	0	0.05	0.01	0.02	0.03
C	0	0.08	0.9	0.01	0	0	0.01
ER	0.04	0	0	0.92	0	0.02	0.02
P	0.01	0.01	0.03	0.03	0.91	0.01	0
S	0.03	0.01	0	0	0	0.94	0.02
V	0.01	0	0	0	0	0.03	0.96

Table 5: Syntaxnet confusion matrix.

	AD	B	C	ER	P	S	V
AD	0.72	0.03	0	0	0.02	0.07	0.01
B	0.03	0.87	0	0.01	0	0.04	0.01
C	0	0.08	0.90	0.01	0.01	0	0
ER	0.06	0	0	0.92	0	0.01	0.01
P	0.01	0	0.02	0.02	0.93	0.01	0
S	0	0	0	0	0	0.97	0.02
V	0.01	0	0	0	0	0.04	0.95

Table 6: DRAGNN confusion matrix.

5 Conclusion

We have presented a comparative performance assessment of five state-of-the-art PoS taggers on the DFA-TIscrivo corpus of K-12 student essays, along with an analysis of the patterns that can be observed in the mistakes made by individual taggers. As this is still a work in progress, the results in the paper are limited to a subset of the corpus containing fifth grade essays. These results provide a valuable baseline that could likely be improved with domain adaptation. On the other hand, it is fair to ask whether the DFA-TIscrivo corpus is different enough from standard Italian to warrant domain adaptation, or whether we would encounter issues with overfitting. In the latter case, an alternative would be the rule-based combination of the output of the five taggers, informed with the knowledge of the observed error patterns.

Acknowledgments

The partial support of SNF through project TIscrivo 2.0 and of SUPSI through project Scripsit is gratefully acknowledged.

References

Abadi et al., 2016 Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Gregory S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat,

Ian J. Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Józefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Gordon Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul A. Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda B. Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *CoRR*, abs/1603.04467.

Abel et al., 2016 Andrea Abel, Aivars Glaznieks, Lionel Nicolas, and Egon Stemle. 2016. An extended version of the koko german L1 learner corpus. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016.*

Alberti et al., 2017 Chris Alberti, Daniel Andor, Ivan Bogatyy, Michael Collins, Dan Gillick, Lingpeng Kong, Terry Koo, Ji Ma, Mark Omernick, Slav Petrov, Chayut Thanapirom, Zora Tung, and David Weiss. 2017. Syntaxnet models for the conll 2017 shared task. *CoRR*, abs/1703.04929.

Andor et al., 2016 Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016. Globally normalized transition-based neural networks. *arXiv preprint arXiv:1603.06042*.

Barbagli et al., 2015 Alessia Barbagli, Piero Lucisano, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2015. Cita: un corpus di produzioni scritte di apprendenti litaliano l1 annotato con errori. In *BProceedings of the Second Italian Conference on Computational Linguistics, CLiC-it*, pages 31–35.

Barbagli et al., 2016 Alessia Barbagli, Pietro Lucisano, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2016. Cita: an l1 italian learners corpus to study the development of writing competence. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, may.

Bohnet and Nivre, 2012 Bernd Bohnet and Joakim Nivre. 2012. A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 1455–1465, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bosco et al., 2016 Cristina Bosco, Fabio Tamburini, Andrea Bolioli, and Alessandro Mazzei. 2016. Overview of the EVALITA 2016 part of speech on twitter for italian task. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016.*

Brunato and Dell'Orletta, 2015 Dominique Brunato and Felice Dell'Orletta. 2015. Isacco: a corpus for investigating spoken and written language development in italian school–age children. *CLiC it*, page 62.

Cignetti and Demartini, 2016 Luca Cignetti and Silvia Demartini. 2016. From data to tools. theoretical and applied problems in the compilation of lissics (the lexicon of written italian in a school context in italian switzerland). *RiCOGNIZIONI. Rivista di Lingue e Letterature straniere e Culture moderne*, 3(6):35–49.

Cignetti et al., 2016 Luca Cignetti, Silvia Demartini, and Simone Fornara. 2016. *Come TIscrivo? La scrittura a scuola tra teoria e didattica*. Aracne.

Cimino and Dell'Orletta, 2016 Andrea Cimino and Felice Dell'Orletta. 2016. Building the state-of-the-art in POS tagging of italian tweets. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016.*

Collins, 2002 Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fornara et al., 2016 Simone Fornara, Luca Cignetti, and Silvia Demartini. 2016. Il lessico di tiscrivo. caratterizzazione del vocabolario e osservazioni in prospettiva didattica. In *Sviluppo della competenza lessicale: acquisizione, apprendimento, insegnamento*, pages 43–60. Aracne, Roma.

Giesbrecht and Evert, 2009 Eugenie Giesbrecht and Stefan Evert. 2009. Is part-of-speech tagging a solved task? an evaluation of pos taggers for the german web as corpus. In I. Alegria, I. Leturia, and S. Sharoff, editors, *Proceedings of the 5th Web as Corpus Workshop (WAC5)*, San Sebastian, Spain.

Kong et al., 2017 Lingpeng Kong, Chris Alberti, Daniel Andor, Ivan Bogatyy, and David Weiss. 2017. DRAGNN: A transition-based framework for dynamically connected neural networks. *CoRR*, abs/1703.04474.

Ling et al., 2015 Wang Ling, Tiago Luís, Luís Marujo, Ramón Fernandez Astudillo, Silvio Amir, Chris Dyer, Alan W Black, and Isabel Trancoso. 2015.

Finding function in form: Compositional character models for open vocabulary word representation. *arXiv preprint arXiv:1508.02096*.

Manning et al., 2014 Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Palmero Aprosio and Moretti, 2016 A. Palmero Aprosio and G. Moretti. 2016. Italy goes to Stanford: a collection of CoreNLP modules for Italian. *ArXiv e-prints*, September.

Ratnaparkhi, 1996 Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Conference on Empirical Methods in Natural Language Processing*, pages 133–142.

Schmid, 1994 Helmut Schmid. 1994. Part-of-speech tagging with neural networks. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 172–176. Association for Computational Linguistics.

Spoustová et al., 2009 Drahomíra "johanka" Spoustová, Jan Hajič, Jan Raab, and Miroslav Spousta. 2009. Semi-supervised training for the averaged perceptron pos tagger. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 763–771, Athens, Greece, March. Association for Computational Linguistics.

Straka et al., 2016 Milan Straka, Jan Hajic, and Jana Straková. 2016. Udpipe: Trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *LREC*.

Straková et al., 2014 Jana Straková, Milan Straka, and Jan Hajič. 2014. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland, June. Association for Computational Linguistics.

syn, 2016 2016. Syntaxnet. `https://www.tensorflow.org/versions/r0.12/tutorials/syntaxnet`. Accessed: 2017-07-13.

Toutanova and Manning, 2000 Kristina Toutanova and Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13*, EMNLP '00, pages 63–70. Association for Computational Linguistics.

Toutanova et al., 2003 Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics.

Zhang and Weiss, 2016 Yuan Zhang and David Weiss. 2016. Stack-propagation: Improved representation learning for syntax. *arXiv preprint arXiv:1603.06598*.