Overview of the EVALITA 2016 Part Of Speech on TWitter for ITAlian Task

Cristina Bosco
Dip. di Informatica, Università di Torino
bosco@di.unito.it

Fabio Tamburini,
FICLIT, University of Bologna, Italy
fabio.tamburini@unibo.it

Andrea Bolioli
CELI
abolioli@celi.it

Alessandro Mazzei
Dip. di Informatica, Università di Torino
mazzei@di.unito.it

Abstract

English. The increasing interest for the extraction of various forms of knowledge from micro-blogs and social media makes crucial the development of resources and tools that can be used for automatically deal with them. PoSTWITA contributes to the advancement of the state-of-the-art for Italian language by: (a) enriching the community with a previously not existing collection of data extracted from Twitter and annotated with grammatical categories, to be used as a benchmark for system evaluation; (b) supporting the adaptation of Part of Speech tagging systems to this particular text domain.

Italiano. *La crescente rilevanza dell'estrazione di varie forme di conoscenza da testi derivanti da micro-blog e social media rende cruciale lo sviluppo di strumenti e risorse per il trattamento automatico. PoSTWITA si propone di contribuire all'avanzamento dello stato dell'arte per la lingua italiana in due modi: (a) fornendo alla comunità una collezione di dati estratti da Twitter ed annotati con le categorie grammaticali, risorsa precedentemente non esistente, da utlizzare come banco di prova nella valutazione di sistemi; (b) promuovendo l'adattamento a questo particolare dominio testuale dei sistemi di Part of Speech tagging che partecipano al task.*

Authors order has been decided by coin toss.

1 Introduction and motivation

In the past the effort on Part-of-Speech (PoS) tagging has mainly focused on texts featured by standard forms and syntax. However, in the last few years the interest in automatic evaluation of social media texts, in particular from microblogging such as Twitter, has grown considerably: the so-called user-generated contents have already been shown to be useful for a variety of applications for identifying trends and upcoming events in various fields.

As social media texts are clearly different from standardized texts, both regarding the nature of lexical items and their distributional properties (short messages, emoticons and mentions, threaded messages, etc.), Natural Language Processing methods need to be adapted for deal with them obtaining reliable results in processing. The basis for such an adaption are tagged social media text corpora (Neunerdt *et al.*, 2013) for training and testing automatic procedures. Even if various attempts to produce such kind of specialised resources and tools are described in literature for other languages (e.g. (Gimpel *et al.*, 2011; Derczynski *et al.*, 2013; Neunerdt *et al.*, 2013; Owoputi *et al.*, 2013)), Italian currently completely lacks of them both.

For all the above mentioned reasons, we proposed a task for EVALITA 2016 concerning the domain adaptation of PoS-taggers to Twitter texts. Participants to the evaluation campaign were required to use the two following data sets provided by the organization to set up their systems: the first one, henceforth referred to as Development Set (DS), contains data manually annotated using a specific tagset (see section 2.2 for the tagset description) and must be used to train participants systems; the second one, referred to as Test Set

(TS), contains the test data in blind format for the evaluation and has been given to participants in the date scheduled for the evaluation.
For better focusing the task on the challenges related to PoS tagging, but also for avoiding the boring problem of disappeared tweets, the distributed version of tweets has been previously tokenised, splitting each token on a different line.
Moreover, according to an "open task" perspective, participants were allowed to use other resources with respect to those released for the task, both for training and to enhance final performances, as long as their results apply the proposed tagsets.

The paper is organized as follows. The next section describes the data exploited in the task, the annotation process and the issues related to the tokenisation and tagging applied to the dataset. The following section is instead devoted to the description of the evaluation metrics and participants results. Finally, we discuss the main issues involved in PoSTWITA.

2 Data Description

For the corpus of the proposed task, we collected tweets being part of the EVALITA2014 SENTIment POLarity Classification (SENTIPOLC) (Basile *et al.*, 2014) task dataset, benefitting of the fact that it is cleaned from repetitions and other possible sources of noise. The SENTIPOLC corpus originates from a set of tweets (Twita) randomly collected (Basile *et al.*, 2013), and a set of posts extracted exploiting specific keywords and hashtags marking political topics (SentiTUT) (Bosco *et al.*, 2013).

In order to work in a perspective of the development of a benchmark where a full pipeline of NLP tools can be applied and tested in the future, the same selection of tweets has been exploited in other EVALITA2016 tasks, in particular in the EVALITA 2016 SENTiment POLarity Classification Task (SENTIPOLC) (Barbieri *et al.*, 2016), Named Entity rEcognition and Linking in Italian Tweets (NEEL-IT) (Basile *et al.*, 2016) and Event Factuality Annotation Task (FactA) (Minard *et al.*, 2016).
Both the development and test set of EVALITA2016 has been manually annotated with PoS tags. The former, which has been distributed as the DS for PoSTWITA, includes 6,438 tweets (114,967 tokens). The latter, that is the TS, is instead composed by 300 tweets (4,759 tokens).

The tokenisation and annotation of all data have been first carried out by automatic tools, with a high error rate which is motivated by the features of the domain and text genre. We adapted the Tweet-NLP tokeniser (Gimpel *et al.*, 2011) to Italian for token segmentation and used the TnT tagger (Brants, 2000) trained on the Universal Dependencies corpus (v1.3) for the first PoS-tagging step (see also section 2.2).

The necessary manual correction has been applied by two different skilled humans working independently on data. The versions produced by them have been compared in order to detect disagreements, conflicts or residual errors which have been finally resolved by the contribution of a third annotator.
Nevertheless, assuming that the datasets of PoSTWITA are developed from scratch for what concerns the tokenisation and annotation of grammatical categories, we expected the possible presence of a few residual errors also after the above described three phases of the annotation process. Therefore, during the evaluation campaign, and before the date scheduled for the evaluation, all participants were invited and encouraged to communicate to the organizers any errors found in the DS. This allowed the organizers (but not the participants) to update and redistribute it to the participants in an enhanced form.

No lexical resource has been distributed with PoSTWITA 2016 data, since each participant is allowed to use any available lexical resource or can freely induce it from the training data.

All the data are provided as plain text files in UNIX format (thus attention must be paid to newline character format), tokenised as described in section 2.1, but only those of the DS have been released with the adequate PoS tags described in section 2.2. The TS contains only the tokenised words but not the correct tags, that have to be added by the participant systems to be submitted for the evaluation. The correct tokenised and tagged data of the TS (called gold standard TS), exploited for the evaluation, has been provided to the participants after the end of the contest, together with their score.

According to the treatment in the dataset from where our data are extracted, each tweet in PoSTWITA corpus is considered as a separate entity and

we did not preserved thread integrity, thus taggers participating to the contest have to process each tweet separately.

2.1 Tokenisation Issues

The problem of text segmentation (tokenisation) is a central issue in PoS-tagger evaluation and comparison. In principle, for practical applications, every system should apply different tokenisation rules leading to different outputs.

We provided in the evaluation campaign all the development and test data in tokenised format, one token per line followed by its tag (when applicable), following the schema:

```
_____ID_TWEET_1_____      _____162545185920778240_____
<TOKEN_1> <TAG1>          Governo PROPN
<TOKEN_2> <TAG2>          Monti PROPN
<TOKEN_3> <TAG3>          :  PUNCT
<TOKEN_4> <TAG4>          decreto NOUN
<TOKEN_5> <TAG5>          in ADP
<TOKEN_6> <TAG6>          cdm PROPN
<TOKEN_7> <TAG7>          per ADP
<TOKEN_8> <TAG8>          approvazione NOUN
<TOKEN_9> <TAG9>          !  PUNCT
<TOKEN_10> <TAG10>        http://t.co/Z76KLLGP URL

_____ID_TWEET_2_____      _____192902763032743936_____
<TOKEN_1> <TAG1>          #Ferrara HASHTAG
<TOKEN_2> <TAG2>          critica VERB
<TOKEN_3> <TAG3>          #Grillo HASHTAG
<TOKEN_4> <TAG4>          perché SCONJ
<TOKEN_n> <TAGn>          ...
```

The first line for each tweet contains the Tweet ID, while the line of each tweet after the last one is empty, in order to separate each post from the following. The example above shows some tokenisation and formatting issues, in particular:

- accents, which are coded using UTF-8 encoding table;
- apostrophe, which is tokenised separately only when used as quotation mark, not when signalling a removed character (like in *dell'/orto*)

All the other features of data annotation are described in details in the following parts of this section.

For what concerns tokenisation and tagging principles in EVALITA2016 PoSTWITA, we decided to follow the strategy proposed in the Universal Dependencies (UD) project for Italian[1] applying only minor changes, which are motivated by the special features of the domain addressed in the task. This makes the EVALITA2016-PoSTWITA gold standard annotation compliant with the other UD datasets, and strongly improves the portability of our newly developed datasets towards this standard.

Assuming, as usual and more suitable in PoS tagging, a neutral perspective with respect to the solution of parsing problems (more relevant in building treebanks), we differentiated our format from that one applied in UD, by maintaining the word unsplitted rather than splitted in different tokens, also in the two following cases:

- for the articulated prepositions (e.g. *dalla* (from-the[fem]), *nell´* (in-the[masc]), *al* (to-the), ...)
- for the clitic clusters, which can be attached to the end of a verb form (e.g. *regalaglielo* (gift-to-him-it), *dandolo* (giving-it), ...)

For this reason, we decided also to define two novel specific tags to be assigned in these cases (see section 1): *ADP_A* and *VERB_CLIT* respectively for articulated prepositions and clitics, according to the strategy assumed in previous EVALITA PoS tagging evaluations.

The participants are requested to return the test file using exactly the same tokenisation format, containing exactly the same number of tokens. The comparison with the reference file will be performed line-by-line, thus a misalignment will produce wrong results.

2.2 Tagset

Beyond the introduction of the novel labels cited above, motivated by tokenisation issues and related to articulated prepositions and clitic clusters, for what concerns PoS tagging labels, further modifications with respect to UD standard are instead motivated by the necessity of more specific labels to represent particular phenomena often occurring in social media texts. We introduced therefore new Twitter-specific tags for cases that following the UD specifications should be all classified into the generic *SYM* (symbol) class, namely emoticons, Internet addresses, email addresses, hashtags and mentions (*EMO*, *URL*, *EMAIL*, *HASHTAG* and *MENTION*). See Table 1 for a complete description of the PoSTWITA tagset.

We report in the following the more challenging issues addressed in the development of our data sets, i.e. the management of proper nouns and of foreign words.

[1] http://universaldependencies.org/it/pos/index.html

Tagset		Category	Examples
UD	**PoSTWITA16**		**if different from UD specs**
ADJ	ADJ	Adjective	-
ADP	ADP	Adposition (simple prep.)	di, a, da, in, con, su, per
	ADP_A	Adposition (prep.+Article)	dalla, nella, sulla, dell
ADV	ADV	Adverb	-
AUX	AUX	Auxiliary Verb	-
CONJ	CONJ	Coordinating Conjunction	-
DET	DET	Determiner	-
INTJ	INTJ	Interjection	-
NOUN	NOUN	Noun	-
NUM	NUM	Numeral	-
PART	PART	Particle	-
PRON	PRON	Pronoun	-
PROPN	PROPN	Proper Noun	-
PUNCT	PUNCT	punctuation	-
SCONJ	SCONJ	Subordinating Conjunction	-
SYM	SYM	Symbol	-
	EMO	Emoticon/Emoji	:-) ^_^ ♥ :P
	URL	Web Address	http://www.somewhere.it
	EMAIL	Email Address	someone@somewhere.com
	HASHTAG	Hashtag	#staisereno
	MENTION	Mention	@someone
VERB	VERB	Verb	-
	VERB_CLIT	Verb + Clitic pronoun cluster	mangiarlo, donarglielo
X	X	Other or RT/rt	-

Table 1: EVALITA2016 - PoSTWITA tagset.

2.2.1 Proper Noun Management

The annotation of named entities (NE) poses a number of relevant problems in tokenisation and PoS tagging. The most coherent way to handle such kind of phenomena is to consider each NE as a unique token assigning to it the PROPN tag. Unfortunately this is not a viable solution for this evaluation task, and, moreover, a lot of useful generalisation on n-gram sequences (e.g. *Ministero/dell/Interno* PROPN/ADP_A/PROPN) would be lost if adopting such kind of solution. Anyway, the annotation of sequences like *Banca Popolare* and *Presidente della Repubblica Italiana* deserve some attention and a clear policy.
Following the approach applied in Evalita 2007 for the PoS tagging task, we annotate as PROPN those words of the NE which are marked by the uppercase letter, like in the following examples:

Banca PROPN	Presidente PROPN	Ordine PROPN
Popolare PROPN	della ADP_A	dei ADP_A
	Repubblica PROPN	Medici PROPN
	Italiana PROPN	

Nevertheless, in some other cases, the uppercase letter has not been considered enough to determine the introduction of a PROPN tag:
"...anche nei Paesi dove..., *"...in contraddizione con lo Stato sociale..."*.
This strategy is devoted to produce a data set that incorporates the speakers linguistic intuition about this kind of structures, regardless of the possibility of formalization of the involved knowledge in automatic processing.

2.2.2 Foreign words

Non-Italian words are annotated, when possible, following the same PoS tagging criteria adopted in UD guidelines for the referring language. For instance, *good-bye* is marked as an interjection with the label INTJ.

3 Evaluation Metrics

The evaluation is performed in a black box approach: only the systems output is evaluated. The evaluation metric will be based on a token-by-

Team ID	Team	Affiliations
EURAC	E.W. Stemle	Inst. for Specialised Commun. and Multilingualism, EURAC Research, Bolzano/Bozen, Italy
ILABS	C. Aliprandi, L De Mattei	Integris Srl, Roma, Italy
ILC-CNR	A. Cimino, F. Dell'Orletta	Istituto di Linguistica Computazionale Antonio Zampolli CNR, Pisa, Italy
MIVOQ	Giulio Paci	Mivoq Srl, Padova, Italy
NITMZ	P. Pakray, G. Majumder	Deptt. of Computer Science & Engg., Nat. Inst. of Tech., Mizoram,Aizawl, India
UniBologna	F. Tamburini	FICLIT, University of Bologna, Italy
UniDuisburg	T. Horsmann, T. Zesch	Language Technology Lab Dept. of Comp. Science and Appl. Cog. Science, Univ. of Duisburg-Essen, Germany
UniGroningen	B. Plank, M. Nissim	University of Groningen, The Nederlands
UniPisa	G. Attardi, M. Simi	Dipartimento di Informatica, Universit di Pisa, Italy

Table 2: Teams participating at the EVALITA2016 - PoSTWITA task.

token comparison and only a single tag is allowed for each token. The considered metric is the Tagging accuracy: it is defined as the number of correct PoS tag assignment divided by the total number of tokens in TS.

4 Teams and Results

16 teams registered for this task, but only 9 submitted a final run for the evaluation. Table 2 outlines participants' main data: 7 participant teams belong to universities or other research centres and the last 2 represent private companies working in the NLP and speech processing fields.

Table 3 describes the main features of the evaluated systems w.r.t. the core methods and the additional resources employed to develop the presented system.

In the Table 4 we report the final results of the PoSTWITA task of the EVALITA2016 evaluation campaign. In the submission of the result, we allow to submit a single "official" result and, optionally, one "unofficial" result ("UnOFF" in the table): UniBologna, UniGroningen, UnPisa and UniDuisburg decided to submit one more unofficial result. The best result has been achieved by the ILC-CNR group (93.19% corresponding to $4,435$ correct tokens over $4,759$).

5 Discussion and Conclusions

Looking at the results we can draw some provisional conclusions about the PoS-tagging of Italian tweets:

- as expected, the performances of the automatic PoS-taggers when annotating tweets are lower than when working on normal texts, but are in line with the state-of-the art for other languages;
- all the top-performing systems are based on Deep Neural Networks and, in particular, on Long Short-Term Memories (LSTM) (Hochreiter, Schmidhuber, 1997; Graves, Schmidhuber, 1997);
- most systems use word or character embeddings as inputs for their systems;
- more or less all the presented systems make use of additional resources or knowledge (morphological analyser, additional tagged corpora and/or large non-annotated twitter corpora).

Looking at the official results, and comparing them with the experiments that the participants devised to set up their own system (not reported here, please look at the participants' reports), it is possible to note the large difference in performances. During the setup phase most systems, among the top-performing ones, obtained coherent results well above 95/96% of accuracy on the development set (either splitting it into a training/validation pair or by making cross-validation tests), while the best performing system in the official evaluation exhibit performances slightly above 93%. It is a huge difference for this kind of task, rarely observed in literature.

One possible reason that could explain this difference in performances regards the kind of docu-

Team ID	Core methods	Resources (other than DS)
EURAC	LSTM NN (word&char embeddings)	DiDi-IT
ILABS	Perceptron algorithm	word features extracted from proprietary resources and 250k entries of wikitionary.
ILC-CNR	two-branch BiLSTM NN (word&char embeddings)	Morhological Analyser (65,500 lemmas) + ItWaK corpus
MIVOQ	Tagger combination based on Yamcha	Evalita2009 Pos-tagged data ISTC pronunciation dictionary
NITMZ	HMM bigram model	-
UniBologna	Stacked BiLSTM NN + CRF (augmented word embeddings)	Morhological Analyser (110,000 lemmas) + 200Mw twitter corpus
UniDuisburg	CRF classifier	400Mw Twitter corpus
UniGroningen	BiLSTM NN (word embedding)	Universal Dependencies v1.3 74 kw tagged Facebook corpus
UniPisa	BiLSTM NN + CRF (word&char embeddings)	423Kw tagged Mixed corpus 141Mw Twitter corpus

Table 3: Systems description.

#	Team ID	Tagging Accuracy
1	ILC-CNR	0.9319 (4435)
2	UniDuisburg	0.9286 (4419)
3	UniBologna_UnOFF	0.9279 (4416)
4	MIVOQ	0.9271 (4412)
5	UniBologna	0.9246 (4400)
6	UniGroningen	0.9225 (4390)
7	UniGroningen_UnOFF	0.9185 (4371)
8	UniPisa	0.9157 (4358)
9	UniPisa_UnOFF	0.9153 (4356)
10	ILABS	0.8790 (4183)
11	NITMZ	0.8596 (4091)
12	UniDuisburg_UnOFF	0.8178 (3892)
13	EURAC	0.7600 (3617)

Table 4: EVALITA2016 - PoSTWITA participants' results with respect to Tagging Accuracy. "UnOFF" marks unofficial results.

ments in the test set. We inherited the development set from the SENTIPOLC task at EVALITA2014 and the test set from SENTIPOLC2016 and, maybe, the two corpora, developed in different epochs and using different criteria, could contain also different kind of documents. Differences in the lexicon, genre, etc. could have affected the training phase of taggers leading to lower results in the evaluation phase.

References

Barbieri, F., Basile, V., Croce, D., Nissim, M., Novielli, N., Patti, V. Overview of the EVALITA 2016 SENTiment POLarity Classification Task. author=, *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)*.

Basile, v., Bolioli, A., Nissim, M., Patti, V., Rosso, P. 2014. Overview of the Evalita 2014 SENTIment POLarity Classification Task *In Proceedings of Evalita 2014*, 50–57.

Basile, P., Caputo, A., Gentile, A.L., Rizzo, G. Overview of the EVALITA 2016 Named Entity rEcognition and Linking in Italian Tweets (NEEL-IT) 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)*.

Basile, V., Nissim, M. Sentiment analysis on Italian tweets. *In Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

Bosco, c., Patti, V., Bolioli, A. Developing Corpora for Sentiment Analysis: The Case of Irony and Senti-TUT. *IEEE Intelligent Systems, special issue on Knowledge-based approaches to content-level sentiment analysis*. Vol 28 num 2.

Brants, T. 2000. TnT – A Statistical Part-of-Speech Tagger. *In Proceedings of the 6th Applied Natural Language Processing Conference*.

Derczynski, L.,Ritter, A., Clark, S., Bontcheva, K. 2013. Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data *In Proceedings of RANLP 2013*, 198–206.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.A. 2011. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. *In Proceedings of ACL 2011*.

Graves, A, Schmidhuber, J. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6), 602–610.

Hochreiter, S., Schmidhuber, J. 1997. Long short-term memory. *Neural Computation*, 9(8), 1735–1780.

Minard, A.L., Speranza, M., Caselli, T. 2016 The EVALITA 2016 Event Factuality Annotation Task (FactA). *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)*.

Neunerdt, M., Trevisan, B., Reyer, M., Mathar, R. 2013. Part-of-speech tagging for social media texts. *Language Processing and Knowledge in the Web*. Springer, 139–150.

Owoputi, O., OConnor, B., Dyer, C., Gimpel, K., Schneider, N., Smith, N.A. 2013. Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters. *In Proceedings of NAACL 2013*.