EVALITA 2016: Overview of the 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian

Pierpaolo Basile
University of Bari
Via E.Orabona, 4
70126 Bari, Italy
basilepp@di.uniba.it

Franco Cutugno
University Federico II
Via Claudio 21
80126 Naples, Italy
cutugno@unina.it

Malvina Nissim
University of Groningen
Oude Kijk in t Jatstraat 26
9700 AS Groningen, NL
m.nissim@rug.nl

Viviana Patti
University of Turin
c.so Svizzera 185
I-10149 Torino, Italy
patti@di.unito.it

Rachele Sprugnoli
FBK and University of Trento
Via Sommarive
38123 Trento, Italy
sprugnoli@fbk.eu

1 Introduction

EVALITA[1] is the evaluation campaign of Natural Language Processing and Speech Tools for the Italian language. The aim of the campaign is to improve and support the development and dissemination of resources and technologies for Italian. Indeed, many shared tasks, covering the analysis of both written and spoken language at various levels of processing, have been proposed within EVALITA since its first edition in 2007. EVALITA is an initiative of the Italian Association for Computational Linguistics[2] (AILC) and it is endorsed by the Italian Association of Speech Science[3] (AISV) and by the NLP Special Interest Group of the Italian Association for Artificial Intelligence[4] (AI*IA).

Following the success of the four previous editions, we organised EVALITA 2016 around a set of six shared tasks and an application challenge. In EVALITA 2016 several novelties were introduced on the basis of the outcome of two questionnaires and of the fruitful discussion that took place during the panel "Raising Interest and Collecting Suggestions on the EVALITA Evaluation Campaign" held in the context of the second Italian Computational Linguistics Conference (CLiC-it 2015) (Sprugnoli et al., 2016). Examples of these novelties are a greater involvement of industrial companies in the organisation of tasks, the introduction of a task and a challenge that are strongly application-oriented, and the creation of cross-task shared data. Also, a strong focus has been placed on using social media data, so as to promote the investigation into the portability and adaptation of existing tools, up to now mostly developed for the newswire domain.

2 Tasks and Challenge

As in previous editions, both the tasks and the final workshop were collectively organised by several researchers from the community working on Italian language resources and technologies. At this year's edition, the community organised six tasks and one additional challenge.

Standard Tasks Of the six tasks organised in the context of this year EVALITA, five dealt with different aspects of processing written language, with a specific focus on social media, and one on speech. A brief description of each task is given below.

- **ArtiPhon** – *Articulatory Phone Recognition*. In this task, participants had to build a speaker-dependent phone recognition system that is to be evaluated on mismatched speech rates. While training data consists of read speech where the speaker was required to keep a constant speech rate, testing data range from slow and hyper-articulated speech to fast and hypo-articulated speech (Badino, 2016).

[1] http://www.evalita.it
[2] http://www.ai-lc.it/
[3] http://www.aisv.it/
[4] http://www.aixia.it/

- **FactA** – *Event Factuality Annotation*. In this task, the factuality profiling of events is represented by means of three attributes associated to event mentions, namely: certainty, time, and polarity. Participating systems were required to provide the values for these three attributes (Minard et al., 2016).

- **NEEL-it** – *Named Entity rEcognition and Linking in Italian Tweets*. The task consists in automatically annotating each named entity mention (belonging to the following categories: Thing, Event, Character, Location, Organization, Person and Product) in a tweet by linking it to the DBpedia knowledge base (Basile et al., 2016).

- **PoSTWITA** – *POS tagging for Italian Social Media Texts*. The task consists in Part-Of-Speech tagging tweets, rather than more standard texts, that are provided in their already tokenised form (Bosco et al., 2016).

- **QA4FAQ** – *Question Answering for Frequently Asked Questions*. The goal of this task is to develop a system retrieving a list of relevant FAQs and corresponding answers related to a query issued by an user (Caputo et al., 2016).

- **SENTIPOLC** – *SENTIment POLarity Classification*. The task consists in automatically annotating tweets with a tuple of boolean values indicating the messages subjectivity, its polarity (positive or negative), and whether it is ironic or not (Barbieri et al., 2016).

Application Challenge In addition to the more standard tasks described above, for the first time EVALITA included a *challenge*, organised by IBM Italy. The **IBM Watson Services Challenge**'s aim is to create the most innovative app on Bluemix services[5], which leverages at least one Watson Service, with a specific focus on NLP and speech services for Italian (`http://www.evalita.it/2016/tasks/ibm-challenge`).

3 Participation

The tasks and the challenge of EVALITA 2016 attracted the interest of a large number of researchers, for a total of 96 single registrations. Overall, 34 teams composed of more than 60 individual participants from 10 different countries[6] submitted their results to one or more different tasks of the campaign.

A breakdown of the figures per task is shown in Table 1. With respect to the 2014 edition, we collected a significantly higher number of registrations (96 registrations *vs* 55 registrations collected in 2014), which can be interpreted as a signal that we succeeded in reaching a wider audience of researchers interested in participating in the campaign. This result could be also be positively affected by the novelties introduced this year to improve the dissemination of information on EVALITA, e.g. the use of social media such as Twitter and Facebook. Also the number of teams that actually submitted their runs increased in 2016 (34 teams *vs* 23 teams participating in the 2014 edition), even if we reported a substantial gap between the number of actual participants and those who registered.

Table 1: Registered and actual participants

task	**registered**	**actual**
ARTIPHON	6	1
FactA	13	0
NEEL-IT	16	5
QA4FAQ	13	3
PoSTWITA	18	9
SENTIPOLC	24	13
IBM Challenge	6	3
total	96	34

In order to better investigate this issue and gather some insights on the reasons of the significant drop in the number of participants w.r.t. the registrations collected, we ran an online questionnaire specifically designed for those who did not submit any run to the task to which they were registered. In two weeks we collected 14 responses which show that the main obstacles to the actual participation in a task were related to personal issues ("I had an unexpected personal or professional problem outside EVALITA" or

[5] `https://console.ng.bluemix.net/catalog/`

[6] Brazil, France, Germany, India, Ireland, Italy, Mexico, The Netherlands, Spain, Switzerland.

"I underestimated the effort needed") or personal choices ("I gave priority to other EVALITA tasks"). As for this last point, NEEL-it and SENTIPOLC were preferred to FactA, which did not have any participant. Another problem mentioned by some of the respondents is that the evaluation period was too short: this issue is highlighted mostly by those who registered to more than one task.

4 Making Cross-task Shared Data

As an innovation at this year's edition, we aimed at creating datasets that would be shared across tasks so as to provide the community with multi-layered annotated data to test end-to-end systems. In this sense, we encouraged task organisers to annotate the same instances, each task with their respective layer. The involved tasks were: SENTIPOLC, PoSTWITA, NEEL-it and FactA.

The testsets for all four tasks comprise exactly the same 301 tweets, although Sentipolc has a larger testset of 2000 tweets, and FactA has an additional non-social media testset of 597 newswire sentences. Moreover, the training sets of PoSTWITA and NEEL-it are almost entirely subsets of SENTIPOLC. 989 tweets from the 1000 that make NEEL-it's training set are in SENTIPOLC, and 6412 of PoSTWITA (out of 6419) also are included in the SENTIPOLC training set.

Table 2: Overview of cross-task shared data. Number of tweets are reported. When the figure is marked with a *, it is instead the number of sentences from newswire documents.

TRAIN

	SENTIPOLC	NEEL-it	PoSTWITA	FactA
SENTIPOLC	7410	989	6412	0
NEEL-it	989	1000	0	0
PoSTWITA	6412	0	6419	0
FactA	0	0	0	2723*

TEST

	SENTIPOLC	NEEL-it	PoSTWITA	FactA
SENTIPOLC	2000	301	301	301
NEEL-it	301	301	301	301
PoSTWITA	301	301	301	301
FactA	301	301	301	597*+301

The matrix in Table 2 shows both the total number of test instances per task (diagonally) as well as the number of overlapping instances for each task pair. Please note that while SENTIPOLC, NEEL-it, and PoSTWITA provided training and test sets made up entirely of tweets, FactA included tweets only in one of their test set, as a pilot task. FactA's training and standard test sets are composed of newswire data, which we report in terms of number of sentences (Minard et al., 2016). For this reason the number of instances in Table 2 is broken down for FactA's test set: 597 newswire sentences and 301 tweets, the latter being the same as the other tasks.

5 Towards Future Editions

On the basis of this edition's experience, we would like to conclude with a couple of observations that prospective organisers might find useful when designing future editions.

Many novelties introduced in EVALITA 2016 proved to be fruitful in terms of cooperation between academic institutions and industrial companies, balance between research and applications, quantity and quality of annotated data provided to the community. In particular, the involvement of representatives from companies in the organisation of tasks, the development of shared data, the presence of application-oriented tasks and challenge are all elements that could be easily proposed also in future EVALITA editions.

Other innovations can be envisaged for the next campaign. For example, in order to help those who want to participate in more than one task, different evaluation windows for different tasks could be planned instead of having the same evaluation deadlines for all. Such kind of flexibility could foster the participation of teams to multiple tasks, but the fact that it impacts on the work load of the EVALITA's organizers should not be underestimated. Moreover, social media texts turned out to be a very attractive

domain but others could be explored as well. For instance, Humanities resulted as one of the most appealing domains in the questionnaires for industrial companies and former participants and other countries are organising evaluation exercises on it (see, for example, the *Translating Historical Text* shared task at CLIN 27[7]).

References

Leonardo Badino. 2016. The ArtiPhon Challenge at Evalita 2016. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *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)*. Associazione Italiana di Linguistica Computazionale (AILC).

Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the EVALITA 2016 SENTiment POLarity Classification Task. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *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)*. Associazione Italiana di Linguistica Computazionale (AILC).

Pierpaolo Basile, Annalina Caputo, Anna Lisa Gentile, and Giuseppe Rizzo. 2016. Overview of the EVALITA 2016 Named Entity rEcognition and Linking in Italian Tweets (NEEL-IT) Task. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *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)*. Associazione Italiana di Linguistica Computazionale (AILC).

Cristina Bosco, Fabio Tamburini, Andrea Bolioli, and Alessandro Mazzei. 2016. Overview of the EVALITA 2016 Part Of Speech on TWitter for ITAlian Task. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *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)*. Associazione Italiana di Linguistica Computazionale (AILC).

Annalina Caputo, Marco de Gemmis, Pasquale Lops, Franco Lovecchio, and Vito Manzari. 2016. Overview of the EVALITA 2016 Question Answering for Frequently Asked Questions (QA4FAQ) Task. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *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)*. Associazione Italiana di Linguistica Computazionale (AILC).

Anne-Lyse Minard, Manuela Speranza, and Tommaso Caselli. 2016. The EVALITA 2016 Event Factuality Annotation Task (FactA). In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *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)*. Associazione Italiana di Linguistica Computazionale (AILC).

Rachele Sprugnoli, Viviana Patti, and Franco Cutugno. 2016. Raising Interest and Collecting Suggestions on the EVALITA Evaluation Campaign. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *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)*. Associazione Italiana di Linguistica Computazionale (AILC).

[7] http://www.ccl.kuleuven.be/CLIN27/