Evalita 2018: Overview on the 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian

Tommaso Caselli
Rijksuniversiteit Groningen
Groningen, The Netherlands
t.caselli@gmail.com

Nicole Novielli
Dipartimento di Informatica
Universit degli Studi di Bari Aldo Moro, Italy
nicole.novielli@uniba.it

Viviana Patti
Dipartimento di Informatica
Universit degli Studi di Torino, Italy
patti@di.unito.it

Paolo Rosso
PRHLT Research Center
Universitat Politcnica de Valncia, Spain
prosso@dsic.upv.es

1 Introduction

EVALITA[1] is the evaluation campaign of Natural Language Processing and Speech Tools for Italian. Since 2007, the general objective of EVALITA is to promote the development and dissemination of language resources and technologies for Italian, providing a shared framework where different systems and approaches can be evaluated in a consistent manner.

EVALITA is an initiative of the Italian Association for Computational Linguistics (AILC)[2] and it is endorsed by the Italian Association for Artificial Intelligence (AI*IA)[3] and the Italian Association for Speech Sciences (AISV)[4].

2 Tasks and Challenge

For the 2018 edition, ten tasks are organized along the following tracks:

Affect, Creativity and Style

- *ABSITA - Aspect-based Sentiment Analysis*. The task is organized as a cascade of two subtasks consisting in automatically annotating sentences from hotel reviews with respect to the identified aspects (Aspect Category Detection (ACD) subtask) and the polarity associated to each one of them (Aspect Category Polarity (ACP) subtask) (Basile et al., 2018a);

- *ITAMoji - Italian Emoji Prediction*. The goal of this task is to develop a system for predicting the most likely emoji associated to a tweet. For simplicity purposes, tweets including only one emoji are considered (Ronzano et al., 2018);

- *IronITA - Irony Detection in Twitter*. The task aims at automatically identifying ironic tweets along two different subtasks. Specifically, the irony detection subtask (Task A) is a binary classification task where the systems are required to predict whether a tweet is ironic or not, while the second subtask (Task B) focuses on the identification of the different types of irony, with special attention to sarcasm recognition (Cignarella et al., 2018);

- *GxG - Cross-Genre Gender Prediction*. This task addresses the problem of gender prediction across different textual genres. Specifically, given a collection of texts from a specific genre, the gender of the author has to be predicted as either female or male. A dataset from different genres is distributed to the participants and gender prediction has to be done either (i) using a model which has been trained on the same genre, or (ii) using a model which has been trained on anything but that genre (Dell'Orletta and Nissim, 2018).

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

Dialogue Systems

- *iLISTEN - itaLIan Speech acT labEliNg*. This task consists in automatically annotating dialogue turns with speech act labels, i.e. with the communicative intention of the speaker, such as statement, request for information, agreement, opinion expression, general answer (Basile and Novielli, 2018).
- *IDIAL - Italian DIALogue systems evaluation*. The task develops and applies evaluation protocols for the quality assessment of dialogue systems for the Italian language. The target of the evaluation are existing task-oriented dialogue systems, both from industry and academia (Cutugno et al., 2018).

Hate Speech

- *AMI - Automatic Misogyny Identification*. This task focuses on the automatic identification of misogynous content both in English and in Italian languages in Twitter. More specifically, it is a two-fold task. It includes: (i) a Misogyny Identification subtask consisting in a binary classification of tweets as being either misogynous or not; (ii) a Misogynistic Behaviour and Target Classification subtask aimed at classifying tweets according to different finer-grained types of misogynistic behaviour detected, such as sexual harassment or discredit, and the target of the message (individuals or group of people). (Fersini et al., 2018a);
- *HaSpeeDe - Hate Speech Detection*. This task is organized into three sub-tasks, concerning: (i) the identification of hate speech on Facebook (HaSpeeDe-FB), (ii) the identification of hate speech on Twitter (HaSpeeDe-TW), and (iii) the cross-dataset setting concerning the assessment of the performance of the hate speech recognition system developed, i.e., when trained on Facebook data and evaluated on Twitter data, and vice versa (Bosco et al., 2018).

Semantics4AI

- *NLP4FUN - Solving language games*. This task consists in designing a solver for "The Guillotine" game, inspired by an Italian TV show. The game involves a single player, who is given a set of five words - the clues - each linked in some way to a specific word that represents the unique solution of the game. Words are unrelated to each other, but each of them has a hidden association with the solution. Once the clues are given, the player has to provide the unique word representing the solution. The participant systems are required to build an artificial player able to solve the game (Basile et al., 2018b).
- *SUGAR - Spoken Utterances Guiding Chef's Assistant Robots*. This task goal is to develop a voice-controlled robotic agent to act as a cooking assistant. To this aim, a train corpus of spoken commands is collected and annotated using a 3D virtual environment that simulates a real kitchen where users can interact with the robot. The task specifically focuses on a set of commands, whose semantics is defined according to the various possible combination of actions, items (i.e. ingredients), tools and different modifiers (Di Maro et al., 2018).

3 Fostering Reproducibility and Cross-community Engagement

Open access to resources and research artifacts, such as data, tools, and dictionaries, is deemed crucial for the advancement of the state of the art in scientific research. Accessibility of resources and experimental protocols enable both full and partial replication of studies in order to further validate their findings, towards building of new knowledge based on solid empirical evidence. To foster reproducibility and encourage follow-up studies leveraging the resources built within EVALITA 2018, we introduced two novelties this year. First of all, we intend to distribute all datasets used as benchmark for the tasks of this edition. To this aim, we have set up a repository on Github[5], in line with the good practices already applied by the organizers of the previous edition[6]. Also, the datasets for all the tasks will be

[5]The dataset of EVALITA 2018 made available by the task organizers can be found at: `https://github.com/evalita2018/data`

[6]The datasets of EVALITA 2016 can be found at: `https://github.com/evalita2016/data`

hosted and distributed by the European Language and Resources Association (ELRA). In addition, we decided to further encourage the sharing of resources by making availability of the systems an eligibility requirement for the best system award (see Section 4).

In the same spirit, we encouraged cross-community involvement in both task organization and participation. We welcomed the initiative of the organizers of AMI, the Automatic Misogyny Identification task (Fersini et al., 2018a), focusing on both English and Italian tweets. This task has been proposed first at IberEval 2018 for Spanish and English (Fersini et al., 2018b), and then re-proposed at Evalita for Italian, and again for English with a new dataset for training and testing. The ITAmoji shared task was also a re-proposal for the Italian language of the *Multilingual Emoji Prediction Task* at International Workshop on Semantic Evaluation (SemEVAL 2018) (Barbieri et al., 2018), which focused on English and Spanish. Here the re-proposal of the task at Evalita was driven by twofold aim to widen the setting for cross-language comparisons for emoji prediction in Twitter and to experiment with novel metrics to better assess the quality of the automatic predictions, also proposing a comparison with human performances on the same task.

In the 2016 edition task organisers were encouraged to collaborate on the creation of a shared test set across tasks (Basile et al., 2017). We were happy to observe that also this year this practice was maintained. In particular, a portion of the dataset of IronITA (Cignarella et al., 2018), the task on irony detection in Twitter, partially overlaps with the dataset of the hate speech detection task (HaSpeeDe) (Bosco et al., 2018). The intersection includes tweets related to three social groups deemed as potential target for hate speech online: immigrants, Muslims and Roma. Also, the sentiment corpora with multi-layer annotations developed in last years by the EVALITA community, which included also morpho-syntactic and entity linking annotations, were exploited by some ABSITA (Basile et al., 2018a) and IronITA (Cignarella et al., 2018) participants to address the finer-grained sentiment related tasks proposed this year under the *Affect, Creativity and Style track*.

4 Award: Best System Across-tasks

For the first time, this year we decided to award the best system across-task, especially that of young researchers. The award was introduced with the aim of fostering student participation to the evaluation campaign and to the workshop, and received a funding from *Google Research*, *CELI*[7], and from the *European Language and Resources Association* (ELRA)[8].

Criteria for eligibility, are (i) the availability of the system as open source software by the end of the evaluation period, when the results are due to the task organizers, and (ii) the presence of at least one PhD candidate, a master or a bachelor student among the authors of the final report describing the system. The systems will be evaluated based on:

- *novelty*, to be declined as novelty of the approach with respect to the state of the art (e.g. a new model or algorithm), or novelty of features (for discrete classifiers);
- *originality*, to be declined as identification of new linguistic resources employed to solve the task (for instance, using WordNet should not be considered as a new resource), or identification of linguistically motivated features; or implementation of theoretical framework grounded in linguistics;
- *critical insight*, to be declined as a deep error analysis that highlights limits of the current system and pave direction to future challenges; technical soundness and methodological rigor.

We collected 7 system nominations from the organizers of 5 tasks belonging to the *Affect, Creativity and Style track* and to the *Hate Speech track*. 14 students were involved in the development of the systems which received a mentions: 7 PhD students and and 7 master students. Most students are enrolled in Italian universities, but 5 of them. The award recipient(s) will be announced during the final EVALITA workshop, co-located with CliC-it 2018, the Fifth Italian Conference on Computational Linguistics[9].

[7] https://www.celi.it/
[8] http://elra.info/en/
[9] http://clic2018.di.unito.it/it/home/

5 Participation

The tasks and the challenge of EVALITA 2018 attracted the interest of a large number of researchers from academia and industry, for a total of 237 single preliminary registrations. Overall, 50 teams composed of 115 individuals from 13 different countries participated to one or more tasks, submitting a total of 34 system descriptions.

Table 1: Registered and actual participants, with overall number of teams and submitted runs.

Track	**Task**	**Participants**		**Teams**	**Submitted runs**
		Registered	*Actual*		
Affect, Creativity, and Style	ABSITA	30	11	7	20
	ITAMOJI	28	11	5	12
	IronITA	30	14	7	24
	GxG	15	9	3	50
Dialogue Systems	iLISTEN	16	4	2	2
	IDIAL	12	12	3	N/A
Hate Speech	AMI	39	16	10	73
	HaSpeeDe	40	32	9	55
Semantics4AI	NLP4FUN	17	4	2	3
	SUGAR	10	2	2	3
Total		237	115	50	242

A breakdown of the figures per task is shown in Table 1. With respect to the 2016 edition, we collected a significantly higher number of both preliminary registrations (237 registrations vs. 96 collected in 2016), teams (50 vs. 34 in 2016), and participants (115 [10] vs. 60 in 2016), that can be interpreted as a signal that we succeeded in reaching a wider audience of researchers interested in participating in the campaign as well as a further indication of the growth of the NLP community at large. This result could be also positively affected by the novelties introduced this year to involve cross-community participation, represented by the ITAMoji and AMI tasks. Indeed, of the 50 teams that submitted at least one run, 12 include researchers from foreign institutions. In addition to this, this year all tasks have received at least one submission.

A further aspect of the success for this edition can be due to the tasks themselves, especially the "Affect, Creativity and Style" and the "Hate Speech" tracks. Although these two tracks cover 60% of all tasks, they have collected the participation of 82% of the teams (41 teams). This is clearly a sign of growing interest in the NLP community at large in the study and analysis of new text types such as those produced in Social Media platforms and (on-line) user-generated content, also reflecting the outcome of the 2016 survey (Sprugnoli et al., 2016).

Finally, we consider the new protocol for the submission of participants' runs, consisting in three non-overlapping evaluation windows, as a further factor that may have positively impact the participation. Indeed, from the 2016 survey, it emerges that the main reasons for not participating in the evaluation either refer to personal issues or preferences ("I gave priority to other EVALITA tasks") also due to the difficulty of participating in the evaluation step of all tasks simultaneously, as the evaluation period was perceived as too short to enable participation to more than one task (Sprugnoli et al., 2016). Although appreciated by the EVALITA participants, this is not a major cause of the increased participation: out of 50 teams, only 6 have participated in more than one task.

Finally, it is compelling to open a reflection on the distinction between constrained and unconstrained submissions and participation to the tasks. Half of the tasks, namely ABSITA, ITAMOji, IronITA, and AMI, paid attention to this distinction and the other half did not take it into account. In early evaluation campaigns, the distinction used to be very relevant as it aimed at distinguishing the contribution of features or the learning approach from external sources of information, mainly intended as lexical

[10] Please note that the unique participants that also submitted a report are 68. This drop is mainly due to the participation to more than one task, resulting in the submission of only one report from the same team.

resources. In recent years, the spread and extensive use of pre-trained word embedding representations, especially as a strategy to initialize Neural Network architectures, challenges this distinction at its very heart. Furthermore, this distinction is also challenged by the development of multi-task learning architectures. A multi-task system could definitely represent an instance of an unconstrained system, although it exploits data from a different task, rather than a lexical resource or additional data annotated with the same information as that in the main task. As a contribution to the discussion on this topic, we think that proponents of tasks that aim at differentiating between constrained and unconstrained runs must specify what are the actual boundaries, in terms of extra training data, auxiliary tasks, use of word embeddings and lexical resources.

6 Final Remarks

For this edition of EVALITA we introduced novelties towards supporting reproducibility and cross-community engagement, towards advancement of methodology and techniques for natural language and speech processing tasks beyond the performance improvement, which is typically used as a metrics to assess state of the art approaches in benchmarking and shared task organization. In particular, the decision to award the best-system across tasks is inspired by this vision and aim at emphasizing the value of critical reflection and insightful discussion beyond the metric-based evaluation of participating systems.

In line with the suggestion provided by the organizers of the previous edition in 2016 (Basile et al., 2016; Sprugnoli et al., 2016), we introduced a novel organization of the evaluation period based on non-overlapping windows, in order to help those who want to participate in more than one task. This year EVALITA has reached a new milestone concerning the participation of industry. Overall, we have registered a total of 9 industrial participants: 7 directly participated to tasks, 6 of them submitted a paper, and 2 were involved as "targets" of an evaluations exercise (Cutugno et al., 2018).

Finally, a new trend that has emerged this year is the presence of tasks, GxG and HaSpeeDe, that aimed at testing the robustness of systems across text genres, further challenging the participants to develop their system. This "extra challenge" aspect is a new trend in EVALITA that started with the 2016 SENTIPOLC task (Barbieri et al., 2016), where the text genre was not changed but the test data was partially created using tweets that do not exactly match the selection procedure used for the creation of the training set.

Acknowledgments

We would like to thank our sponsors CELI[11], Google Research and the European Language and Resources Association (ELRA)[12] for their support to the event and to the best-system across task award. A further thank goes to ELRA for its offer and support in hosting the task datasets and systems' results. We also thanks Agenzia per l'Italia Digitale (AGID)[13] for its endorsement.

References

Francesco Barbieri, Basile Valerio, Croce Danilo, Nissim Malvina, Novielli Nicole, and Patti Viviana. 2016. Overview of the Evalita 2016 SENTIment POLarity Classification Task. In Pierpaolo Basile, Franco Cutugno, Malvina Nissim, Viviana Patti, and Rachele Sprugnoli, editors, *Proceedings of the 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2016)*, Turin, Italy. CEUR.org.

Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. 2018. Semeval 2018 task 2: Multilingual emoji prediction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 24–33. Association for Computational Linguistics.

Pierpaolo Basile and Nicole Novielli. 2018. Overview of the Evalita 2018 itaLIan Speech acT labEliNg (iLISTEN) Task. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the 6th*

[11] https://www.celi.it/
[12] http://elra.info/en/
[13] https://www.agid.gov.it

evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18), Turin, Italy. CEUR.org.

Pierpaolo Basile, Franco Cutugno, Malvina Nissim, Viviana Patti, and Rachele Sprugnoli. 2016. EVALITA 2016: Overview of the 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. In Pierpaolo Basile, Franco Cutugno, Malvina Nissim, Viviana Patti, and Rachele Sprugnoli, editors, *Proceedings of the 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2016)*, Turin, Italy. CEUR.org.

Pierpaolo Basile, Malvina Nissim, Rachele Sprugnoli, Viviana Patti, and Francesco Cutugno. 2017. Evalita goes social: Tasks, data, and community at the 2016 edition. *Italian Journal of Computational Linguistics*, 3(1).

Pierpaolo Basile, Valerio Basile, Danilo Croce, and Marco Polignano. 2018a. Overview of the EVALITA 2018 Aspect-based Sentiment Analysis task (ABSITA). In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, Turin, Italy. CEUR.org.

Pierpaolo Basile, Marco de Gemmis, Lucia Siciliani, and Giovanni Semeraro. 2018b. Overview of the EVALITA 2018 Solving language games (NLP4FUN) Task. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)* , Turin, Italy. CEUR.org.

Cristina Bosco, Felice Dell'Orletta, Fabio Poletto, Manuela Sanguinetti, and Maurizio Tesconi. 2018. Overview of the EVALITA 2018 Hate Speech Detection Task. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)* , Turin, Italy. CEUR.org.

Alessandra Teresa Cignarella, Simona Frenda, Valerio Basile, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2018. Overview of the EVALITA 2018 Task on Irony Detection in Italian Tweets (IronITA). In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)* , Turin, Italy. CEUR.org.

Francesco Cutugno, Maria Di Maro, Sara Falcone, Marco Guerini, Bernardo Magnini, and Antonio Origlia. 2018. Overview of the EVALITA 2018 Evaluation of Italian DIALogue systems (IDIAL) Task. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)* , Turin, Italy. CEUR.org.

Felice Dell'Orletta and Malvina Nissim. 2018. Overview of the EVALITA 2018 Cross-Genre Gender Prediction (GxG) Task. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)* , Turin, Italy. CEUR.org.

Maria Di Maro, Antonio Origlia, and Francesco Cutugno. 2018. Overview of the EVALITA 2018 Spoken Utterances Guiding Chef's Assistant Robots (SUGAR) Task. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)* , Turin, Italy. CEUR.org.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018a. Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI). In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)* , Turin, Italy. CEUR.org.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018b. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In Paolo Rosso, Julio Gonzalo, Raquel Martínez, Soto Montalvo, and Jorge Carrillo de Albornoz, editors, *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018.*, volume 2150 of *CEUR Workshop Proceedings*, pages 214–228. CEUR-WS.org.

Francesco Ronzano, Francesco Barbieri, Endang Wahyu Pamungkas, Viviana Patti, and Francesca Chiusaroli. 2018. Overview of the EVALITA 2018 Italian Emoji Prediction (ITAMoji) Task. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)* , Turin, Italy. CEUR.org.

Rachele Sprugnoli, Viviana Patti, and Cutugno Franco. 2016. Raising Interest and Collecting Suggestions on the EVALITA Evaluation Campaign. In Pierpaolo Basile, Franco Cutugno, Malvina Nissim, Viviana Patti, and Rachele Sprugnoli, editors, *Proceedings of the 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2016)*, Turin, Italy. CEUR.org.