Chapter 19
Italian EVALITA Benchmark Linguistic Resources, NLP Services and Tools

Viviana Patti, Valerio Basile, Andrea Bolioli, Alessio Bosca, Cristina Bosco, Michael Fell, and Rossella Varvara

Abstract Starting from the first edition held in 2007, EVALITA is the initiative for the evaluation of Natural Language Processing tools for Italian. We describe the EVALITA4ELG project, whose main aim is to systematically collect the resources released as benchmarks for this evaluation campaign, and make them easily accessible through the European Language Grid platform. The collection is moreover integrated with systems and baselines as a pool of web services with a common interface, deployed on a dedicated hardware infrastructure.

1 Overview and Objectives of the Pilot Project

In Natural Language Processing (NLP), periodic campaigns are a popular means to set benchmarks for specific tasks, stimulate the development of comparable systems and ultimately promote research advancement (Nissim et al. 2017). The validation of NLP models on different datasets strongly depends on the possibility of generalising their results on data and languages other than those on which they have been trained and tested (Magnini et al. 2008). Recent trends are pushing towards proposing benchmarks for multiple tasks (Wang et al. 2018), or for testing the adaptability of systems to different textual domains, genres, and languages, including under-researched and under-resourced ones. The recent specific emphasis on multilingual assessment is also driven by a growing awareness that language technologies can help promote multilingualism and linguistic diversity (Joshi et al. 2020). In this context, the EVALITA4ELG project integrates linguistic resources and language technologies developed under the umbrella of the EVALITA evaluation campaign into the European Language Grid.

Viviana Patti · Valerio Basile · Cristina Bosco · Michael Fell · Rossella Varvara
University of Turin, Italy, viviana.patti@unito.it, valerio.basile@unito.it, cristina.bosco@unito.it, michael.fell@unito.it, rosella.varvara@unito.it

Andrea Bolioli · Alessio Bosca
CELI, Italy, andrea.bolioli@h-farm.com, alessio.bosca@h-farm.com

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_19

EVALITA[1] is an initiative of the Italian Association for Computational Linguistics (Associazione Italiana di Linguistica Computazionale, AILC[2]). Since 2007, it has been providing a shared framework where different systems and approaches can be evaluated and compared with each other with respect to a large variety of tasks, organised by the Italian research community. The focus of EVALITA is to support the advancement of methodologies and techniques for natural language and speech processing in an historical perspective, beyond the performance improvement, favouring reproducibility and cross-community engagement.

The main goal of the EVALITA4ELG project is to leverage more than a decade of findings of the Italian NLP community, in order to provide easier access to resources and tools for Italian through ELG. We worked towards the achievement of multiples goals, namely: (i) a survey of the tasks organised in the seven editions of EVALITA, released as a knowledge graph; (ii) an anonymisation procedure for improving compliance with current data standard policies; (iii) the integration of resources and systems developed during EVALITA into the ELG platform; (iv) the creation of a unified benchmark for evaluating Italian Natural Language Understanding (NLU); (v) the dissemination of a shared protocol and a set of best practices to describe new resources and tasks in a format that allows a quick integration of metadata into the European Language Grid.

2 Methodology

We started by surveying the tasks organised in EVALITA, collecting the resources and their metadata for upload, and organising this set of information in an ontology. We anonymised the resources according to the current policies for the protection of people's privacy. Finally, we integrated systems and baselines as a pool of web services with a common interface.

2.1 Surveying the EVALITA Tasks

Starting in 2007, EVALITA has been devoted to the evaluation of NLP tools for Italian, providing a shared framework in which participating systems are evaluated on a growing set of different tasks. Rather than being focused on a single task, EVALITA has always been characterised by a wider variety of tasks: each edition of the EVALITA campaign, held in 2007 (Magnini et al. 2008), 2009, 2011 (Magnini et al. 2013), 2014 (Attardi et al. 2015), 2016 (P. Basile et al. 2017), 2018 (Caselli et al. 2018) and 2020 (V. Basile et al. 2020), has been organised around a set of shared tasks dealing with both written and spoken language, varying with respect to the

[1] http://www.evalita.it

[2] https://www.ai-lc.it

challenges tackled and datasets used. The number of tasks has considerably grown, from five tasks, in the first edition in 2007, to 14 tasks in the latest edition held in 2020. Following the trends of other national and international evaluation campaigns, like, e. g., SemEval[3], the typology of tasks also evolved, progressively including a larger variety of exercises oriented to semantics and pragmatics. In particular, the 2016 edition brought a focus on social media data and on the use of shared data across tasks. Open access to resources and research artifacts is deemed crucial for the advancement of the state of the art (Caselli et al. 2018) and the availability of shared evaluation benchmarks is crucial for fostering reproducibility and comparability of results. Organisers were encouraged to collaborate, stimulated to the creation of a shared test set across tasks, and to eventually share all resources with a wider audience. This has resulted in the creation of GitHub public repositories.[4]

2.2 The EVALITA Knowledge Graph

Starting from the semi-structured repositories mentioned in the previous section and from the information collected by surveying seven editions of EVALITA, we built a knowledge graph (KG) that provides the essential information about the editions of the EVALITA evaluation campaign. The KG describes EVALITA in terms of organised tasks, but also of people and institutions that constitute the EVALITA community throughout the years. The KG is structured around an ontology implemented in OWL and it is available both on the website of the EVALITA4ELG project[5] and as a service on the ELG platform. The current version of the ontology comprises 148 classes, 37 object properties and nine data properties. The ontology and the KG are thoroughly described in Patti et al. (2020). As an example, Figure 1 depicts the structure of the KG around the HaSpeeDe2018 task.

The knowledge graph can be queried through a SPARQL endpoint, which allows to inspect the ontology by selecting some variables that occur among the set of triples (subject, predicate, object) composing the knowledge graph. It is thus possible to answer relevant questions related to the EVALITA campaign, extracting information from the KG such as, e. g., “What is the total number of institutions involved as organisers of tasks in all seven EVALITA campaigns?”:

```
SELECT  (COUNT(distinct ?institution) AS ?totalInstitutions)
where {
  ?task e4e:hasInstitution ?institution.
}
>>>> result: 55 <<<<
```

[3] https://semeval.github.io

[4] https://github.com/evalita2016/data

[5] http://evalita4elg.di.unito.it

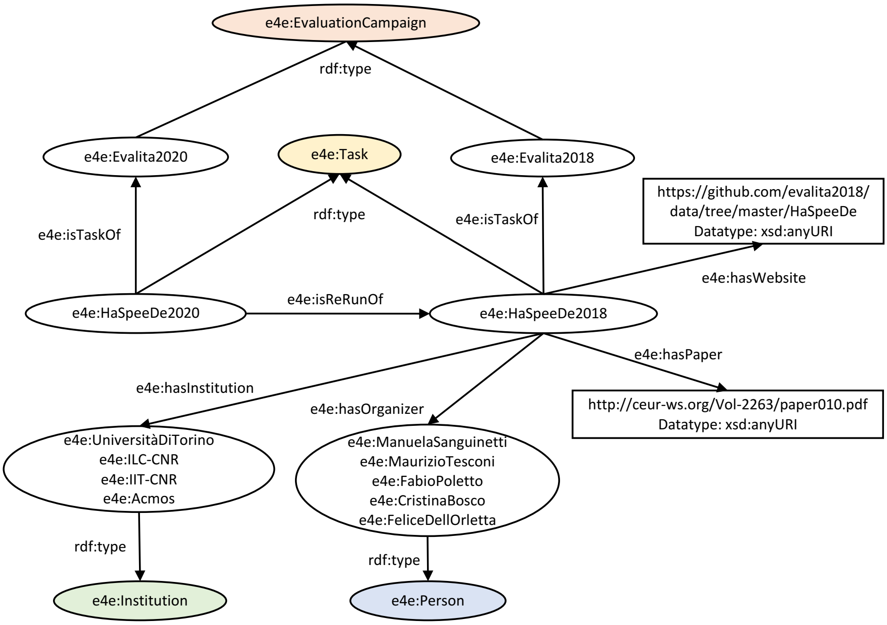

Fig. 1 EVALITA knowledge graph; primary classes are colored and their relations illustrated around the HaSpeeDe2018 task

2.3 Anonymisation of Resources

The EVALITA resources to be made accessible in the ELG platform had to be carefully checked and made compliant with the current policies about data releasing and sharing (e. g., GDPR, Rangel and Rosso 2018), therefore particular attention has been paid to data anonymisation. The datasets collected for EVALITA4ELG were anonymised relying on an automatic anonymisation tool developed in the context of the AnonymAI research project, and then manually reviewed in order to assess their quality. AnonymAI is a nine months research project co-financed by the H2020 project NGI Trust focusing on providing legally compliant anonymisation profiles customised to the needs of end users.

The anonymisation profile applied to the EVALITA4ELG dataset detects and masks person names, phone numbers, email addresses, mentions/replies/retweets, and URLs. The most frequent entities that were masked in the anonymisation process consist of person names and mentions (e. g., in the SardiStance dataset about 50 person names and 150 mentions).

2.4 Release of Data and Models through ELG

At the time of this writing, 51 Language Resources and Technologies are linked to the EVALITA4ELG project in ELG.[6] Eight services were fully integrated into ELG: four of them from the EVALITA 2018 edition, and four of them from the most recent EVALITA 2020 edition. Of the 2018 systems, three are hate speech detection systems (HaSpeeDe 2018 task) and one is Gender Detection (GxG). Of the 2020 systems, two are hate speech detectors (HaSpeeDe 2020 task), one is a POS tagger for spoken language (KIPoS task), and one is a misogyny detection system (AMI task). All datasets and services are accessible interactively from the ELG website or programmatically by means of REST API calls or the ELG-provided Python SDK.

3 Conclusions and Results of the Pilot Project

EVALITA4ELG has been a successful effort towards the inclusion of resources for the Italian language in the European Language Grid. We created a catalogue of resources and models developed during the various editions of the EVALITA campaign, designed in the form of a knowledge graph that can be inspected through SPARQL queries. We collected the original distribution of the resources used for EVALITA tasks and we created 44 entries. For 13 resources, together with CELI, we developed and applied an anonymisation procedure to mask personal and sensitive data. We integrated eight available systems from different tasks into ELG. Finally, we organised an event on September 2021 with hybrid participation[7], including an overview of the project and the results obtained, a tutorial about integrating systems and resources on ELG, and a round table with 14 invited speakers chosen among the most active organisers of tasks of EVALITA.

Acknowledgements The work described in this article has received funding from the EU project European Language Grid as one of its pilot projects. The work has also received funding from the EU's Horizon 2020 research and innovation programme under grant agreement no. 825618 (AnonymAI, NGI Trust).

References

Attardi, Giuseppe, Valerio Basile, Cristina Bosco, Tommaso Caselli, Felice Dell'Orletta, Simonetta Montemagni, Viviana Patti, Maria Simi, and Rachele Sprugnoli (2015). "State of the Art Language Technologies for Italian: The EVALITA 2014 Perspective". In: *Intelligenza Artificiale* 9, pp. 43–61.

[6] https://live.european-language-grid.eu/catalogue/project/1397

[7] http://evalita4elg.di.unito.it/conference

Basile, Pierpaolo, Malvina Nissim, Rachele Sprugnoli, Viviana Patti, and Francesco Cutugno (2017). "EVALITA Goes Social: Tasks, Data, and Community at the 2016 Edition". In: *Italian Journal of Computational Linguistics* 3.1, pp. 93–127.

Basile, Valerio, Danilo Croce, Maria Di Maro, and Lucia C. Passaro (2020). "EVALITA 2020: Overview of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian". In: *Proc. of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), 17 Dec. 2020*. Vol. 2765. CEUR Workshop Proceedings.

Caselli, Tommaso, Nicole Novielli, Viviana Patti, and Paolo Rosso (2018). "Evalita 2018: Overview on the 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian". In: *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. Ed. by Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso. Torino: CEUR Workshop Proceedings, pp. 3–8.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury (2020). "The State and Fate of Linguistic Diversity and Inclusion in the NLP World". In: *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL, pp. 6282–6293.

Magnini, Bernardo, Amedeo Cappelli, Fabio Tamburini, Cristina Bosco, Alessandro Mazzei, Vincenzo Lombardo, Francesca Bertagna, Nicoletta Calzolari, Antonio Toral, Valentina Bartalesi Lenzi, Rachele Sprugnoli, and Manuela Speranza (2008). "Evaluation of Natural Language Tools for Italian: EVALITA 2007". In: *Proc. of the 6th Int. Conference on Language Resources and Evaluation (LREC 2008)*. Marrakech: ELRA, pp. 2536–2543.

Magnini, Bernardo, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, eds. (2013). *Evaluation of Natural Language and Speech Tools for Italian, International Workshop, EVALITA 2011, Rome, Italy, January 24-25, 2012, Revised Selected Papers*. Vol. 7689. Lecture Notes in Computer Science. Springer. URL: https://doi.org/10.1007/978-3-642-35828-9.

Nissim, Malvina, Lasha Abzianidze, Kilian Evang, Rob van der Goot, Hessel Haagsma, Barbara Plank, and Martijn Wieling (2017). "Last Words: Sharing Is Caring: The Future of Shared Tasks". In: *Computational Linguistics* 43.4, pp. 897–904.

Patti, Viviana, Valerio Basile, Cristina Bosco, Rossella Varvara, Michael Fell, Andrea Bolioli, and Alessio Bosca (2020). "EVALITA4ELG: Italian Benchmark Linguistic Resources, NLP Services and Tools for the ELG Platform". In: *Italian Journal of Computational Linguistics* 6.6-2, pp. 105–129. DOI: https://doi.org/10.4000/ijcol.754.

Rangel, Francisco and Paolo Rosso (2018). "On the Implications of the General Data Protection Regulation on the Organisation of Evaluation Tasks". In: *Language and Law* 5.2, pp. 95–117.

Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman (2018). "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding". In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels: ACL, pp. 353–355.