Project abbreviation: EVALITA4ELG

Project name: Italian EVALITA Benchmark Linguistic Resources, NLP Services and Tools for the European Language Grid

Project coordinator: Viviana Patti

Project consortium: University of Turin; CELI

Funding: ELG

Project duration: 1 year

Main key words: Language Resources, EVALITA, Evaluation, Italian, Corpora

Background of the research topic: Starting from the first edition held in 2007, EVALITA9 has been proposed as the initiative devoted to the evaluation of Natural Language Processing tools for Italian, providing a shared framework where participating systems had the possibility to be evaluated on a growing set of different tasks. EVALITA is a biennal initiative of AILC (Associazione Italiana di Linguistica Computazionale), and since 2016 has been co-located with CLiC-it, the Italian Conference on Computational Linguistics. The number of tasks has considerably grown, from 5 tasks, in the first edition in 2007, to 10 tasks in 2018, and 14 tasks in the edition held in 2020 (Basile, Croce, et al. 2020), showing the peculiar vitality of the research community behind this campaign, which involves scholars from academy and industry, from Italy and foreign countries alike. Since their systematic collection and sharing are among the goals of EVALITA from the beginning, by providing a platform for making resources and models more accessible the EVALITA4ELG project represents a meaningful improvement of the evaluation campaign.

Goal of the project: The goal of EVALITA4ELG is to enable the ELG users to access the resources and models for the Italian language produced over the years in the context of the EVALITA evaluation campaign. More precisely, we worked towards the following goals: (i) a survey of the past 62 tasks organized in the seven editions of EVALITA, released as a knowledge graph; (ii) a common anonymization procedure of the resource data for improving their compliance with current 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) systems; (v) the dissemination of a shared protocol and a set of best practices to describe also new resources and new tasks in a format that allows a quick integration of metadata into the ELG platform.

Project abstract: The Italian language is underrepresented in the European Language Grid platform, currently including few Language Technology services and corpora - mostly parallel corpora and multilingual dependency treebanks, focused on texts featured by standard forms and syntax. Our aim is to build the catalogue of EVALITA resources and tasks ranging from traditional tasks like POS-tagging and parsing to recent and popular ones such as sentiment analysis and hate speech detection on social media, and integrate them in the ELG platform. The project includes the integration of state-of-the-art LT services into the ELG platform, accessible as web services.

Publications:

Viviana Patti, 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. Italian Journal of Computational Linguistics 6-2