IJCoL
Italian Journal of Computational Linguistics

6-2 | 2020
Further Topics Emerging at the Sixth Italian Conference on Computational Linguistics

Lessons Learned from EVALITA 2020 and Thirteen Years of Evaluation of Italian Language Technology

Lucia C. Passaro, Maria Di Maro, Valerio Basile and Danilo Croce

OpenEdition Journals

Electronic version
URL: https://journals.openedition.org/ijcol/740
DOI: 10.4000/ijcol.740
ISSN: 2499-4553

Publisher
Accademia University Press

Printed version
Number of pages: 79-102

Brought to you by Università di Pisa - Coordinamento Sistema Bibliotecario

Università di Pisa

Electronic reference
Lucia C. Passaro, Maria Di Maro, Valerio Basile and Danilo Croce, "Lessons Learned from EVALITA 2020 and Thirteen Years of Evaluation of Italian Language Technology", *IJCoL* [Online], 6-2 | 2020, Online since 01 December 2020, connection on 12 November 2021. URL: http://journals.openedition.org/ijcol/740 ; DOI: https://doi.org/10.4000/ijcol.740

Lessons Learned from EVALITA 2020 and Thirteen Years of Evaluation of Italian Language Technology

Lucia C. Passaro*
Università di Pisa

Maria Di Maro**
Università di Napoli "Federico II"

Valerio Basile†
Università di Torino

Danilo Croce‡
Università di Roma "Tor Vergata"

This paper provides a summary of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA2020) which was held online on December 17th, due to the 2020 COVID-19 pandemic. The 2020 edition of Evalita included 14 different tasks belonging to five research areas, namely: (i) Affect, Hate, and Stance, (ii) Creativity and Style, (iii) New Challenges in Long-standing Tasks, (iv) Semantics and Multimodality, (v) Time and Diachrony. This paper provides a description of the tasks and the key findings from the analysis of participant outcomes. Moreover, it provides a detailed analysis of the participants and task organizers which demonstrates the growing interest with respect to this campaign. Finally, a detailed analysis of the evaluation of tasks across the past seven editions is provided; this allows to assess how the research carried out by the Italian community dealing with Computational Linguistics has evolved in terms of popular tasks and paradigms during the last 13 years.

1. Introduction

The Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA) is the biennial initiative aimed at promoting the development of language and speech technologies for the Italian language. Today, EVALITA is organized by the Italian Association of Computational Linguistics (AILC)[1] and endorsed by the Italian Association for Artificial Intelligence (AIxIA)[2] and the Italian Association for Speech Sciences (AISV)[3].

EVALITA provides a shared framework where different systems and approaches can be scientifically evaluated and compared with each other with respect to a wide array of tasks, suggested and organized by the Italian Natural Language Processing (NLP) community. The proposed tasks represent scientific challenges where methods,

* Dept. of Philology, Literature and Linguistics - Via S. Maria 36, 56126, Pisa, Italy.
E-mail: lucia.passaro@fileli.unipi.it
** Via Tarsia, 31, 80135 Naples, Italy E-mail: maria.dimaro2@unina.it
† Dept. of Computer Science - Corso Svizzera, 185, 10149, Turin, Italy.
E-mail: valerio.basile@unito.it
‡ Dept. of Enterprise Engineering - Via del Politecnico 1, 00133 Rome, Italy.
E-mail: croce@info.uniroma2.it
1 http://www.ai-lc.it
2 http://www.aixia.it
3 http://www.aisv.it

resources, and systems can be tested against shared benchmarks representing linguistic open issues or real world applications, possibly in a multilingual and/or multi-modal perspective. The collected data sets represent big opportunities for scientists to explore old and new problems concerning NLP in Italian as well as to develop solutions and to discuss the NLP-related issues within the community. The 2020 edition has seen the organization of both traditionally present tasks and completely new ones.

Over the course of 13 years, EVALITA has seen the organization of several tasks along several topics and research areas. In fact, the challenges proposed to the EVALITA participants in the event evolved according to several aspects including the new trends in international research in Natural Language Processing as well as the release of new instruments (e.g., new types of word embeddings) and datasets (e.g., data from social media) which have posed new challenging tasks and benchmarks. The growth of EVALITA can be seen also by numeric data, with 5 tasks organized in 2007 and 14 tasks organized in 2020), 21 participating teams in 2007 and 51 in 2020, 30 participants in 2007 and 130 in 2020.

This paper is organized as follows. First, an overview of the topics addressed in the EVALITA campaign, considering all the editions organized between 2007 and 2020, is provided. Secondly, several aspects of the 2020 edition are analyzed in detail, including the tasks organized around common thematic areas and the results obtained (Section 3), the participation (Section 4) and the award attributed to the best system across tasks (Section 5). Finally, Section 6 is left for discussing the evolution of the campaign and to draw some conclusions.

2. EVALITA challenges across editions: from 2007 to 2020

As anticipated before, EVALITA is a periodic evaluation campaign on Natural Language Processing and Speech Technologies for the Italian language. More specifically, the campaign focuses on tasks involving written language or speech transcriptions. However, no speech-related tasks were organized after the 2016 edition.

In addition, the challenges presented at EVALITA evolved greatly since its very first edition, in 2007. The changes have been dictated in part by international research trends, and in part by the fact that for some tasks or problems the level of accuracy reached by more modern systems made them less interesting from a research perspective as they can be considered almost solved.

Several aspects of the evaluation campaign have been addressed in previous research work, highlighting both the state of the art from a technological point of view at different historical moments (Attardi et al. 2015; Basile et al. 2017) and the continuous growth of the Italian EVALITA community (Basile et al. 2017). This paper aims at going deeper from a thematic perspective, through an overview of the research topics approached over time. Along these lines, the research interests of the Italian NLP community will be explored, as well as how they have been influenced by the literature in Natural Language Processing, also in the international arena.

The number of tasks proposed has steadily grown over the years, with very few exceptions that still managed to obtain great participation. From 5 challenges proposed in 2007, we reached 14 in the latest edition of 2020, that span a wide variety of topics and areas of interest.

EVALITA tasks were not always grouped by the organizers around common themes, but over the years, and as the number of tasks grew, this grouping became more common. However, given the fact that the organization of the campaign has been carried out independently year by year, the groups of tasks thus became very

heterogeneous from a diachronic point of view. For this reason, we tried to group all the proposed tasks through a manual process, by identifying *post hoc* a common topic. The aim was to group together themes, research areas and tasks that are strictly related to each other or that have similar end goals. The purpose was to better show the evolution of such topics and research areas over the years both from a quantitative point of view (i.e. how many tasks per research area) and of their stability throughout the years. Specifically, we grouped the task topics as follows:

Affect. The tasks concerning the recognition of subjective and emotive traits of texts. Almost all the competitions were organized on datasets collected from social media. The topic first emerged in 2014, with the organization of a task on Sentiment Analysis and a task on Emotion recognition. The list of the tasks belonging to this group are: Sentiment Polarity Classification (2014); Emotion Recogniton Task (2014); SENTIPOLC - SENTIment POLarity Classification (2016); ABSITA - Aspect-based Sentiment Analysis (2018); ITAMoji - Italian Emoji Prediction (2018) IronITA - Irony Detection in Twitter (2018); AMI - Automatic Misogyny Identification (2018); HaSpeeDe - Hate Speech Detection (2018); ATE_ABSITA - Aspect Term Extraction and Aspect-Based Sentiment Analysis (2020); AMI - Automatic Misogyny Identification (2020); SardiStance - Stance Detection (2020); HaSpeeDe - Hate Speech Detection (2020).

Coreference. Tasks on coreference resolution, namely the identification of mentions in the same or in different documents. The topic is strongly correlated to the problem of Named Entity Recognition. Both tasks in this group were organized in 2011, and consist of Cross-document Coreference Resolution (2011) and Anaphora Resolution (2011).

Dialog. Tasks concerning dialogic data. The topic emerged relatively recently, as the interest for automatic dialog system has grown over the last few years and consists of QA4FAQ - Question Answering for Frequently Asked Questions (2016), iLISTEN - itaLIan Speech acT labEliNg (2018) and IDIAL - Italian DIALogue systems evaluation (2018).

Lemmatisation. The group contains the single task named as the group, organized in 2011. Being very often considered as a subproblem of PoS-tagging, such a topic is strongly associated and often solved according to it.

Multimodality. This topic emerged in 2020, as multi-modal systems that encompass both NLP and other disciplines such as Computer Vision become widely researched also internationally. In the case of EVALITA, the DANKMEMES - Multimodal Artefacts Recognition (2020) was organized in the 2020 edition.

NER, Events and Time. Longstanding topic at EVALITA concerning the automatic extraction and classification of temporal expressions and named entities. The problem has been faced on several types of documents and from several points of view. The group includes: Temporal Expression Recognition and Normalization (2007); Named Entity Recognition (2007); Entity Recognition (2009); Named Entity Recognition on Transcribed Broadcast News (2011); Evaluation of Events and Temporal Information (2014); FactA - Event Factuality Annotation (2016); NEEL-IT - Named Entity rEcognition and Linking in Italian Tweets (2016).

Parsing. Tasks concerning syntactic aspects of the texts. The tasks organized in this context include both Parsing and closely related problems such as domain adaptation. The group includes: Parsing (2007); Parsing (2009); Parsing (2011); Domain Adaptation for Dependency Parsing (2011); Dependency Parsing (2014).

PoS Tagging. Tasks related to Part of Speech tagging. As for Named Entity Recognition area, also in this case the tasks have been organized on several types of texts, including social media and transcriptions from spoken especially in the last few years. To this group belong: Part of Speech Tagging (2007); PoS-Tagging (2009); PoSTWITA - POS tagging for Italian Social Media Texts (2016); KIPoS - Part-of-speech Tagging on Spoken Language (2020).

Semantics. All the tasks have been proposed in the past two EVALITA editions (2018 and 2020). We can identify two trends, namely facing specific problems of semantics, and the automatic resolution of language games. This area includes: NLP4FUN - Solving language games (2018); CONcreTEXT - Concreteness in Context (2020); Ghigliottin-AI - Evaluating Artificial Players for the Language Game "La Ghigliottina" (2020) and PRELEARN - Prerequisite Relation Learning (2020).

Senses and Frames. Tasks related to this area have been proposed in the earliest editions of EVALITA. Later years have seen the introduction of Neural Language Models, and thus also the proposed tasks reflected this shift in perspective. The tasks organized in this area are: Word Sense Disambiguation (2007), Lexical Substitution (2009), Super Sense Tagging (2011), Frame Labeling over Italian Texts (2011).

Speech. Problems linked to the area of *speech* are longstanding in the field of NLP. In fact, tasks related to speech have been proposed in most editions of EVALITA. However, the interest for this area in terms of proposed tasks has reduced in the last few years, with no tasks present in the 2020 edition. Proposed tasks were focused on several distinct areas, including evaluation of dialogue systems, different forms of automatic speech recognition and application of speech technologies. The following tasks have been proposed through the years: Connected Digits Recognition (2009), Spoken Dialogue Systems Evaluation (2009), Speaker Identity Verification (2009), Automatic Speech Recognition - Large Vocabulary Transcription (2011), Forced Alignment on Spontaneous Speech (2011) Voice Applications on Mobile (2011), Forced Alignment on Children Speech (2014), Human and Machine Dialect Identification from Natural Speech and Artificial Stimuli (2014), Speech Activity Detection and Speaker Localization in Domestic Environments (2014), ArtiPhon - Articulatory Phone Recognition (2016), SUGAR - Spoken Utterances Guiding chef's Assistant Robots (2018).

Style. Emerging tasks related to the writing style of the texts. The topic emerged in 2018 and includes: GxG - Cross-Genre Gender Prediction (2018); CHANGE-IT - Style Transfer (2020); TAG-it - Topic, Age and Gender Prediction (2020); AcCompl-it-Acceptability & Complexity evaluation (2020).

Textual Entailment. This area includes only the task Textual Entailment of 2009. The problem of determining inferential relations between two portions of text is being revised nowadays in light of the available resources to encode sentence meaning.

Time and Diacrony. The topic concerns problems related to the shift of meaning or the writing style over years. To this group belong two tasks organized in the 2020 edition of EVALITA: DaDoEval - Dating Documents (2020) and DIACR-Ita - Diachronic Lexical Semantics (2020).

Figure 1 shows the number of tasks organized for each edition of EVALITA and the number of different areas studied that year. As shown in the figure, trends for both the number of tasks and research areas are positive. However, we observe a drop specifically for 2014 and 2016. The past two editions were instead very prolific in terms of tasks, with 2020 being the edition with the most overall proposed tasks (14).

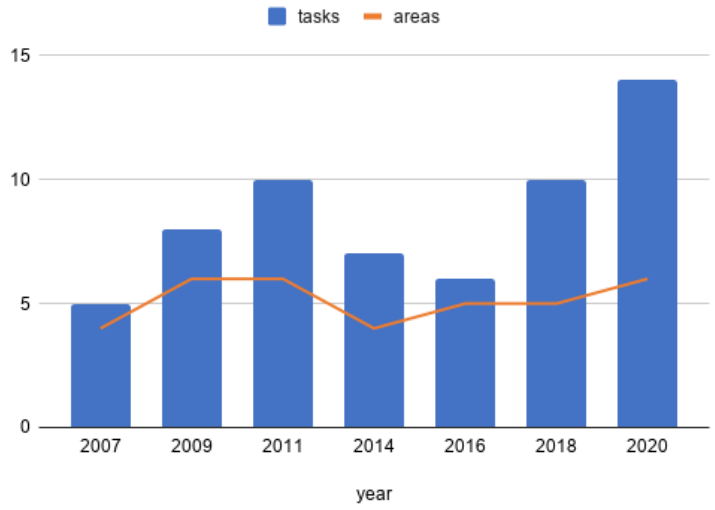

Figure 1
Number of tasks organized in the context of the EVALITA campaign over years. We classified the organized tasks according to common main topics, or research areas. The orange line shows the distribution of areas across years.

In addition to this analysis, we tried to evaluate how tasks and topics evolved over the years.

Figure 2 shows the evolution of each research area across the years. The size of the areas is proportional to the number of tasks proposed for each of them.

In the first edition of EVALITA, most of the proposed tasks concerned standard linguistic annotation problems such as PoS Tagging and Parsing. Moreover, classical information extraction tasks such as Named Entity Recognition were organized as well. These are arguably fundamental building blocks for any further NLP-related task. The development of resources and toolkits to face them was definitely needed at the time. For the next edition in 2011, the speech domain was taken into account as well. Since then, speech-related tasks have remained a stable topic for almost all the subsequent editions, with the interest gradually fading only in the last few years. No tasks in the speech domain were in fact proposed in 2020.

A crucial turning point for EVALITA was 2014. This edition saw the introduction of tasks that are more focused on the application of NLP tools to various problems, such as for example Affect, which were prominently featured in later years. This year also marked a gradual reduction in interest for traditional NLP tasks, in favor of these new aspects and domains. We can identify two factors that may have played an important role in this shift in perspective. On the one hand, advancements in NLP algorithms and toolkits allowed for an easier and more effective resolution of traditional NLP problems. On the other hand, the introduction of mainstream algorithms for language modeling (e.g., word2vec, (Mikolov et al. 2013a, 2013b)) stimulated the interest in more semantic-related tasks which have remained stable in subsequent years.

In addition, 2014 also paved the way for a more marked interest in social media. In fact, a wide array of tasks have been organized since then considering data collected

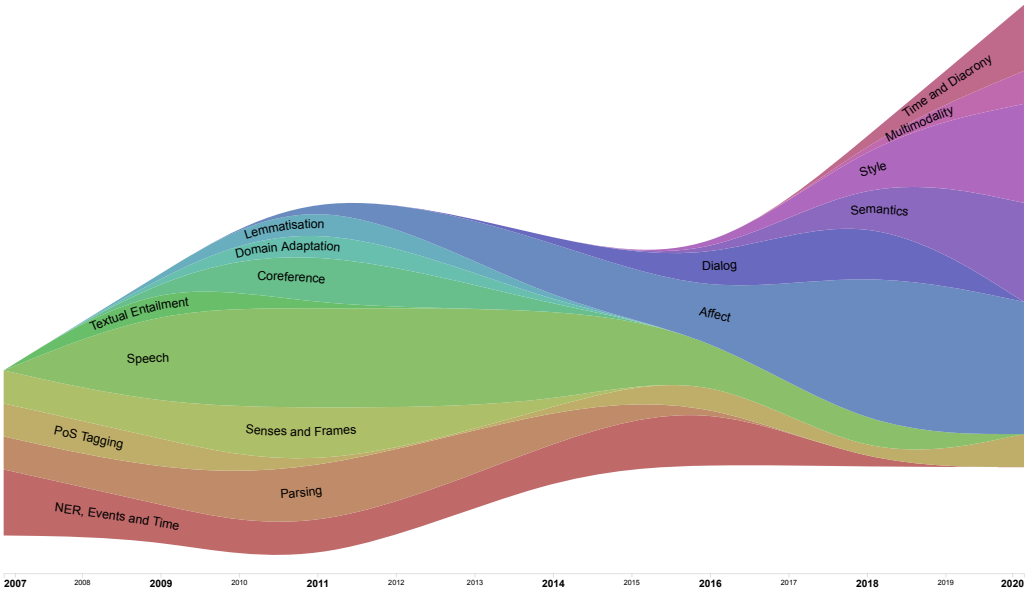

Figure 2
The figure represents the evolution of each research areas across the years. The size is proportional to the number of tasks proposed for each area. Between the first edition in 2007 and the latest one in 2020, some topics registered a loss of interest, while others have emerged.

from them. Such tasks are not exclusively related to a more applicative domain (e.g., affect), but also on facing traditional tasks such as Named Entity recognition revised in the more challenging setting of user-generated texts in social media. This trend has been in line with international research, which has seen an important growth in the analysis of social media data from several perspectives.

Another important milestone for EVALITA was 2018. In this edition in fact, we saw both the introduction of new tasks related to the area of semantics (e.g., for solving language games), and the implementation of systems based on novel language models based on the contextualized representation of words and sentences such as BERT (Devlin et al. 2019). This is also in line with the global research trends that have seen a marked increase in approaches based on these architectures, that both showed state-of-the-art performances in most natural language tasks and allowed for easier modeling of downstream NLP problems thanks to the paradigm of pre-training and fine-tuning.

Finally, we believe that it is interesting to point out the growth of participation over time, particularly marked in the last two editions. This is clearly shown in Figure 3. Indeed, participation has been steadily growing both in terms of teams presenting a system and in terms of actual participants, with 2020 being the most participated year. This is in line with the number of proposed tasks for the various editions (see Figure 1) and also with the growth of the Italian NLP community.

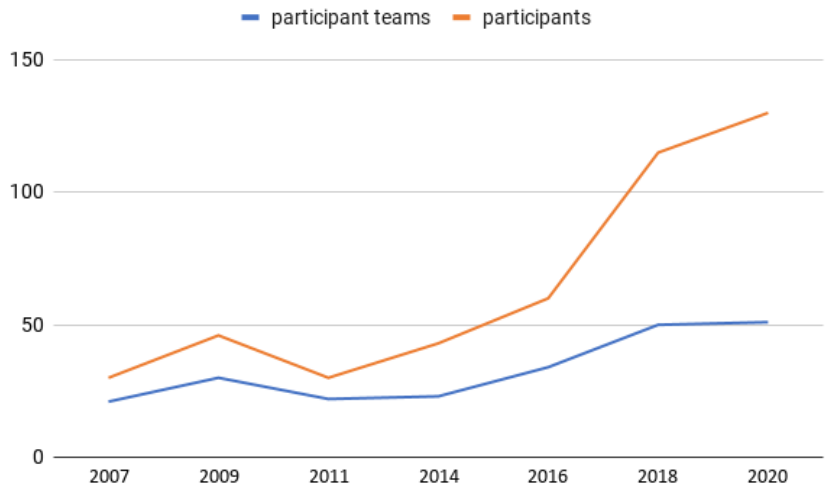

Figure 3
Participation to the EVALITA evaluation campaign from 2007 to 2020. The number of participants has been almost always growing, in line with the growth of the NLP community in Italy.

3. EVALITA 2020 Tracks and tasks

The EVALITA 2020 edition (Basile et al. 2020), held online on December 17^{th} due to the 2020 COVID-19 pandemic, counts 14 different tasks organized along five research areas (tracks) according to their objective and characteristics. The reports of both organizers and participants were peer-reviewed and published on CEUR Workshop Proceedings (CEUR-WS.org, ISSN 1613-0073). Data produced by the task organizers and made available to the participants have been collected on GitHub, with the purpose of making them available in accordance to the terms and conditions of the respective data sources.

The edition has been highly participated, with 51 groups whose participants have affiliation in 14 countries. Although EVALITA is generally promoted and targeted to the Italian research community, this edition saw an international interest and participation, probably also due to the fact that several Italian researchers who contributed to the organization of the tasks or participated in them as authors work in different countries. A snapshot of the geographical origin of EVALITA participants (i.e., their affiliation) is depicted in Section 4 and Figure 4.

The tasks proposed at EVALITA 2020 are grouped according to five main research areas, which are more coarse-grained with respect to the research areas proposed in the previous pages. Such research areas, used as a narrative thread in the EVALITA 2020 proceedings are: (i) *Affect, Hate, and Stance*, (ii) *Creativity and Style*, (iii) *New Challenges in Long-standing Tasks*, (iv) *Semantics and Multimodality*, (v) *Time and Diachrony*.

The tasks organized for each of the proposed research areas as well as the main findings and results obtained for each task are described below.

Affect, Hate, and Stance

The track includes affect-related tasks, mostly concerning social media texts. The proposed tasks in this area are focused on detecting hateful or misogynistic contents as well as aspect-based sentiment analysis or stance.

AMI - Automatic Misogyny Identification (Fersini, Nozza, and Rosso 2020). This shared task is aimed at automatically identifying misogynous contents in Twitter for the Italian language. In particular, the AMI challenge is focused on: (1) recognizing misogynous and aggressive messages and (2) discriminating misogynistic contents from the non-misogynistic ones, while guaranteeing the fairness of the model. This task is one of the most participated, with more than 20 runs in the first sub-task. Participants proposed a wide plethora of methods, ranging from shallow models (e.g., based on logistic regression) to deep neural models, involving Deep Convolution Neural networks and Transformer-based architectures. Results suggest that, while the identification of misogynous text can be considered a more accessible task, the recognition of aggressiveness needs to be properly addressed. Most of the systems obtaining the highest results are based on the straightforward application of Transformer-based architectures, pre-trained on Italian corpora or multilingual ones. In some cases, teams experimented with the use of additional lexical resources such as misogynous and sentiment lexicons. Finally, it is worth noting the portability of some systems: four systems were applied to this task and the HaSpeeDe task, described hereafter.

ATE_ABSITA - Aspect Term Extraction and Aspect-Based Sentiment Analysis (De Mattei et al. 2020b). A task on Aspect Term Extraction (ATE) and Aspect-Based Sentiment Analysis (ABSA). The task is approached as a cascade of three subtasks: Aspect Term Extraction (ATE), Aspect-based Sentiment Analysis (ABSA) and Sentiment Analysis (SA). This subtask provides an extension with respect to the previous one involving Aspect-based Sentiment Analysis, proposed in (Basile et al. 2018): in particular, it focuses on the detection and classification of specific terms expressing the writers' opinions. Three research teams participated from academy and industry. In general, systems proposed the adoption and combination of Transformer-based models pre-trained on Italian corpora or multi-lingual ones.

HaSpeeDe - Hate Speech Detection (Sanguinetti et al. 2020). The task is a rerun of the shared task on hate speech detection at the message level on Italian social media texts proposed for the first time in 2018. The main task is a binary hate speech detection problem, one in-domain and one out-of-domain. On the same data provided for the main task, the topics of stereotypes in communication and nominal utterances are investigated by two pilot tasks. HaSpeeDe has been the most participated one, with 14 participants and about 30 runs submitted in the first sub-task. Again, participants proposed a wide plethora of methods, ranging from shallow models (e.g., based on Support Vector Machine) to deep neural models, mostly involving Transformer-based architectures. Most of the systems obtaining the highest results are based on the straightforward application of Transformer-based architectures, pre-trained on Italian corpora or multilingual ones. Many participants argued the possibility of taking into account the correlation between texts containing hate speech and texts expressing stereotyped ideas about targets, with interesting results. Unfortunately, no groups proposed solutions to the third subtask. In several cases, teams experimented with the use of additional lexical resources to support the recognition of hate. Finally, it is worth noting the porta-

bility of some systems: four systems were applied to this task and the AMI task, described above.

SardiStance - Stance Detection (Cignarella et al. 2020). The goal of the task is to detect the stance of the author towards the target "Sardines movement" in Italian tweets. Two subtasks model (A) Textual Stance Detection and (B) Contextual Stance Detection. Both the subtasks consist of a three-class (in favour, against, neutral) classification problem based only on textual information or on the text enriched with additional information about the author and the post of the tweet. Participants proposed a heterogeneous set of methods, ranging from shallow models (e.g., based on Support Vector Machine) to deep neural models such as Recurrent Neural Networks, Deep Convolutional Networks or Transformer-based architectures. The adoption of Transformers again provided the best results, especially when combined with data augmentation techniques: best results were obtained when the training material is extended with additional examples gathered using distant supervision. Generally, both subtasks were quite challenging, due to the complexity of the phenomenon addressed in the stance detection (i.e., the Sardines movement) and the poor information expressed in very short messages, which are often ironic or sarcastic. As a result, several participants were not able to improve the strong baselines proposed by the organizers, namely a Support Vector Machine for subtask A and a Logistic regression for subtask B, relying on a shallow representation based on simple lexical features.

Creativity and Style

The track involves tasks related to the writing style expressed in texts, both from a detection and a generation perspective.

CHANGE-IT - Style Transfer (De Mattei et al. 2020a). The first natural language generation task for Italian. Change-IT focuses on style transfer performed on the headlines of two Italian newspapers at opposite ends of the political spectrum. Specifically, the goal is to "translate" the headlines from a style to another. This is the very first task focused on natural language generation included in the EVALITA evaluation framework. Although this task did not receive any submission, the organizers release to the participating teams not only training material but also a sequence to a sequence baseline model that performs the task to help everyone to get started with the task.

TAG-it - Topic, Age and Gender Prediction (Cimino, Dell'Orletta, and Nissim 2020). TAG-it is a profiling task for Italian. This task represents a sort of evolution of the Cross-Genre Gender Prediction (GxG) Task (Dell'Orletta and Nissim 2018) presented at EVALITA 2018, whose objective was the automatic detection of a writer's gender given a set of her/his messages. In particular, the task is articulated in two separate subtasks, namely: (i) a first subtask where participants are required to provide the topic which mainly characterizes a set of texts while characterizing their corresponding author through their gender and age; (ii) a second subtask, where input texts are already enriched with information about the underlying topic while systems are required to separately determine the author's age or gender. Teams proposed solutions based on traditional machine learning approaches or deep neural architectures. As in almost all other EVALITA tasks, those approaches adopting Transformer-based architectures obtained the best results. In general,

results suggest that topic and gender recognition tasks are easier to predict than authors' age.

Semantics and Multimodality

This track collects a series of shared tasks involving natural language processing at the level of word and sentence meaning. Some of these tasks go even beyond that, stimulating the development of models that learn and integrate knowledge from structured resources, commonsense knowledge bases, and images.

CONcreTEXT - Concreteness in Context (Gregori et al. 2020). The task focuses on automatic assignment of concreteness values to words in context for the Italian and English languages. Participants are required to develop systems able to rate the concreteness of a target word in a sentence on a scale from 1 (for fully abstract) to 5 (for maximally concrete). The task is structurally similar to word sense disambiguation, whereas the input consists of a word marked in a sentence, and the systems have to predict a numeric value on a 7-point scale indicating its abstractness (or concreteness). Participants were invited to employ external resources of all kinds, e.g., knowledge bases, and indeed several of them did so. In fact, the top-ranked system is a hybrid of neural architectures and knowledge extracted from databases such as behavioral norms.

DANKMEMES - Multimodal Artefacts Recognition (Miliani et al. 2020). This is the first multimodal task for Italian, where participants were asked to create models for several classification tasks on Italian Internet memes about the 2019 Italian Government Crisis, namely Meme Detection, Hate Speech Identification, and Event Clustering. The task was mainly approached with multi-task learning frameworks, leveraging the multimodality, and therefore showcasing similarities and differences between the natural language and the image domains. Interestingly, data augmentation was also used by several systems, in both areas, producing robust classification systems, in particular for the first two subtasks.

Ghigliottin-AI - Evaluating Artificial Players for the Language Game "La Ghigliottina" (Basile et al. 2020). The task challenges researchers to develop a system able to defeat human players at the language game "La Ghigliottina", a highly popular and appreciated quiz game in Italian television. The task is a re-run of the homonym 2018 task also at EVALITA. The system proposed by the task organizers was the only one to sufficiently solve this hard task, as for the 2018 edition as well. Such a solution is based on handcrafted patterns based on domain expertise, whereas the competitor system, based on word embeddings, performed poorly in comparison.

PRELEARN - Prerequisite Relation Learning (Alzetta et al. 2020). The task is devoted to automatically inferring prerequisite relations from educational texts collected from the Italian Wikipedia. The task consists of classifying prerequisite relations between pairs of concepts distinguishing between prerequisite pairs and non-prerequisite pairs. In this task, a "concept" is identified by a Wikipedia page, and the abstract of its page is provided as training material. Variants of the main task are proposed to explore the impact of external resources (*constrained* vs. *unconstrained* tasks) and domain (*in-domain* vs. *cross-domain*). The prevalent approach was neural, with recent transformer-based models achieving the best performance.

Time and Diachrony

This track comprises two tasks that deal with time-related aspects of natural language, namely document dating and semantic shift over time.

DaDoEval - Dating Documents (Menini et al. 2020). The task focuses on assigning a temporal span to a document, by recognising when a document was issued. The task is cast as a classification problem at increasing levels of granularity: coarse-grained (one out of five historical periods), fine-grained (5-year temporal slices), and exact year of publication. The genre dimension is also explored with a subtask variant in a cross-genre setting. The two participant systems both employ supervised machine learning. However, contrary to what is common in most text classification tasks in recent NLP, the system implementing a Support Vector Machine with word- and character-based features outperformed the other system implementing a BERT-based transformer neural network. The latter performed relatively better in the cross-genre subtask

DIACR-Ita - Diachronic Lexical Semantics (Basile et al. 2020). This is the first edition of a shared task on automatic detection of lexical and semantic shift for Italian. The participants were asked to develop systems that can automatically detect if a given word has changed its meaning over time. The input is composed of words in their context, and the task is cast as a classification problem. A wide variety of approaches have been proposed, mostly (but not exclusively) based on word embeddings and several strategies to align such representations. The post-alignment approach, where pre-trained word embeddings are aligned after their individual training, showed the most promising performance on this task.

New Challenges in Long-standing Tasks

One of the goals of EVALITA 2020 as a whole was monitoring the state of the art of language technologies on the Italian language over time. For this reason, the initial call for task proposals included an explicit suggestion to propose "old" (possibly revisited) shared tasks, in order to measure the performance of new models on longstanding open problems in a fast-paced NLP scenario. This unofficial invitation attracted the tasks grouped in this track, which consists of new takes on part-of-speech tagging and acceptability/complexity classification, carried out on new kinds of data and in novel contexts.

AcCompl-it- Acceptability & Complexity evaluation (Brunato et al. 2020). The task is aimed at developing and evaluating methods to classify Italian sentences according to Acceptability and Complexity. A dataset was compiled from several sources, including several datasets in Italian from the Universal Dependencies initiative, and annotated by means of crowdsourcing. The task is formally a regression problem where, given a sentence, participant systems are expected to produce numeric values on a scale from 1 to 7. The two participant systems are radically different, i.e., a rule-based pipeline including several layers of syntactic and semantic analysis, and a transformer-based neural approach. The latter outperformed the former by a large margin. However, the structure of the shared task unveiled interesting properties of the rule-base system with respect to its robustness against synthetic data.

KIPoS - Part-of-speech Tagging on Spoken Language (Bosco et al. 2020). This is a part-of-speech tagging task, with the novel characteristic of being carried out on tran-

scriptions of spoken language, as opposed to written language. The task proposes three subtasks exploring the impact of the speech register (*formal* vs. *informal*) on the classification. The participants employed a mixture of approaches ranging from "classic" POS-tagging models based on Hidden Markov Models and rule-based systems, to recent transformer-based neural architectures. The best system participating in the main task employed a pre-trained BERT-derived neural language model. Domain adaptation techniques and external resources have also been employed in both the main task (ranking second) and the two *cross* subtasks (ranking first).

4. Participation to EVALITA 2020

EVALITA 2020 attracted the interest of a large number of researchers from academia and industry, for a total of 51 teams composed of about 130 individuals participating in one or more of the 14 proposed tasks. After the evaluation period, 58 system descriptions were submitted, i.e., a 70% percent increase with respect to the previous EVALITA edition (Caselli et al. 2018). The peculiarity of this edition also relies on the participants' countries of origin, as the 47,9% came from a non-Italian country (26,8% from European countries, 21,1% from non-European countries). This could depend on the higher accessibility of the event in its virtual form. The origin-related data are displayed in figure 4.

Moreover, task organizers allowed participants to submit more than one system result (called runs), for a total of 240 submitted runs. Table 1 shows the different tracks and tasks along with the number of participating teams and submitted runs. The data reported in the table is based on information provided by the task organizers at the end of the evaluation process. Such data represents an overestimation with respect to the systems described in the proceedings. The trends are similar, but there are differences due to groups participating in more than a task, and groups that have not produced a system report.

Table 1
Number of participating teams and number of runs organized by track and task. The data reported is an overestimation with respect to the systems described in the proceedings (e.g., teams participating in more than a task are counted according to the number of tasks they participated in).

TRACK	TASK	TEAMS	RUNS
Affect, Hate, and Stance	AMI	8	31
	ATE_ABSITA	3	8
	HaSpeeDe	14	27
	SardiStance	12	36
Creativity and Style	CHANGE-IT	0	0
	TAG-it	3	20
New Challenges in Long-standing Tasks	AcCompl-it	2	6
	KIPoS	3	14
Semantics and Multimodality	CONcreTEXT	4	15
	DANKMEMES	5	15
	Ghigliottin-AI	2	2
	PRELEARN	3	14
Time and Diachrony	DaDoEval	2	16
	DIACR-Ita	9	36

Figure 4
Affiliation of the researchers participating at EVALITA 2020. The over 180 proceedings authors (participants and task organizers), have affiliation in 18 countries. The map was built with the *My Maps* service by Google Maps.

Differently from the previous EVALITA editions, the organizers were discouraged from distinguishing the submissions between unconstrained and constrained runs[4]. The rationale for this decision is that the recent 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. Participation was quite imbalanced across different tracks and tasks, as reported in Figure 5: each rectangle represents a task whose size reflects the number of participants, while the color indicated the corresponding track.

In line with the previous editions of EVALITA, the track *Affect, Creativity and Style* covers about half of the total in terms of participating teams. On the one hand, this demonstrates the well-known interest of the NLP community for Social Media platforms and user-generated content. On the other hand, we report a better balance with respect to the 2018 edition, were about 80% of the teams participated in similar tracks (*Affect, Creativity and style* and *Hate Speech*, which have been merged in this edition).

4 A system is considered *constrained* when using the provided training data only; on the contrary, it is considered *unconstrained* when using additional material to augment the training dataset or to acquire additional resources.

Another significant number of teams participated to the *Semantics and Multimodality* and *Time and Diachrony* tracks, while the other tracks where less participated. Unfortunately, no team participated to the *CHANGE-IT* task, mainly due to the complexity of the task.

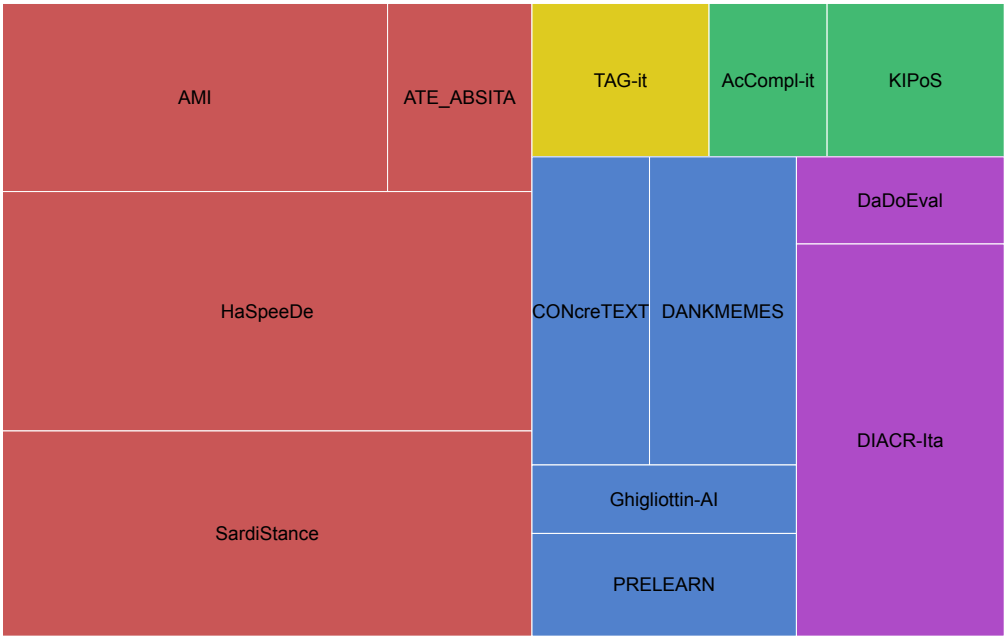

Figure 5
Number of participating teams organized by track (color) and task. The red color is adopted for the track *Affect, Hate, and Stance*, the yellow color for *Creativity and Style*, green for *New Challenges in Long-standing Tasks*, blue for *Semantics and Multimodality* and purple for *Time and Diachrony*.

In addition to being widely participated, the over 180 proceedings authors, including both participants and task organizers, have affiliation in 18 countries, with the 64% from Italy and the 36% of participants from Institutions and companies abroad. The group of the 59 task organizers have affiliations in 6 countries (90% from Italy while 10% from institutions and companies abroad). The gender distribution is highly balanced, with 30 females and males.

5. Award: Best System Across Tasks

In line with the previous edition, EVALITA 2020 confirmed the award to the best system across-task. The award was introduced with the goal of fostering student participation to the evaluation campaign and to the workshop. EVALITA received sponsorship funding from Amazon Science, Bnova s.r.l., CELI s.r.l., the European Language Resources Association (ELRA) and Google Research.

A committee of 5 members was asked to choose the best system across tasks. Four of the five members come from academia and one from the industry. The composition of the committee is balanced with respect to the level of seniority as well as for their academic background (computer science-oriented vs. humanities-oriented): Giuseppe Attardi (University of Pisa, Italy), Giuseppe Castellucci (Amazon, Seattle,

USA), Francesca Chiusaroli (University of Macerata, Italy), Gloria Gagliardi (University of Naples "L'Orientale", Italy), and Nicole Novielli (University of Bari "Aldo Moro", Italy). In order to select a short list of candidates, the task organizers were invited to propose up to two candidate systems participating to their tasks (not necessarily top ranking). The committee was provided with the list of candidate systems and the criteria for eligibility, based on:

- *novelty* with respect to the state of the art;
- *originality*, in terms of identification of new linguistic resources, identification of linguistically motivated features, and implementation of a theoretical framework grounded in linguistics;
- *critical insight*, paving the way to future challenges (deep error analysis, discussion on the limits of the proposed system, discussion of the inherent challenges of the task);
- *technical soundness* and *methodological rigor*.

We collected 10 system nominations from the organizers of 7 tasks from across all tracks. The candidate systems are authored by 20 authors, among whom 12 are students, either at the masters or PhD level. The award recipient(s) were announced during the final EVALITA workshop, during the plenary session, held online.

Two special mentions were given to two systems, namely *rmassidda@DaDoEval* (Massidda 2020), in which the dating classification task was tackled by comparing different models to generate sentence embeddings (USE, LaBSE and SBERT) along with bag-of-entities representations, and *UmBERTo-MTSA@AcCompl-It* (Sarti 2020), tackling the syntactic complexity task with a multi-task training approach with the addition of unlabelled datasets extracted from available Italian treebanks. The systems were commented as follows:

> *rmassidda@DaDoEval: Document Dating Using Sentence Embeddings at EVALITA 2020*: This is a work by a single author who is a student and who deserves encouragement. The paper is sound, the approach is original and we appreciate the intuition of leveraging bag-of-entities to enhance the classification performance. Finally, the system achieved a good performance in a challenging and less popular task.

> *UmBERTo-MTSA@AcCompl-It: Improving Complexity and Acceptability Prediction with Multi-task Learning on Self-Supervised Annotations*: This is a work by a single author who is a student and who deserves encouragement. The paper is sound and well-written; the proposed approach achieved a good performance in a challenging and less popular task. Moreover, the author provided an in-depth analysis of the model and the correlation with linguistic phenomena paving also the way to future directions.

UNITOR@Sardistance2020 (Giorgioni et al. 2020) was instead awarded as the best system with the following motivation:

> *UNITOR@Sardistance2020: Combining Transformer-based architectures and Transfer Learning for robust Stance Detection*: We would like to give the best paper to the UNITOR system as we found a solid and well-written paper proposing a methodologically sound approach for the Sardistance task. In particular, we appreciated the idea of using linguistically motivated auxiliary tasks in the training of the model as well as the usage of Distant Supervision to augment the training dataset.

This approach is original, well-grounded in deep knowledge of the linguistic phenomena under study, and proved effective, as demonstrated by the final ranking in the challenge.

The awarded system, participating in the sub-task A of SardiStance (Cignarella et al. 2020), used a Transformer-based architecture, namely UmBERTo[5], combined with Transfer Learning techniques employing auxiliary tasks, such as Sentiment Detection, Hate Speech Detection and Irony Detection. The participants also used additional datasets, as mentioned in the motivations.

6. Final Remarks

Since 2007, EVALITA has seen the organization of several tasks along with several topics and research areas. Over the years, participants were challenged with both traditional and new tasks according to international research trends in Natural Language Processing. We tried to attribute the proposed tasks over years with a label indicating their topic. From the analysis emerged an overall growing of the campaign in terms of organized shared tasks, participation and topics.

As for the last edition of EVALITA, a record number of 14 tasks were organized. In this edition, the topics of Affect and Semantics were confirmed as two of the most interesting and thriving ones, both in the number of organized tasks and actual participants. In any case, almost all tasks involved the analysis of written texts. In fact, although the KIPoS task considered transcriptions of spoken Italian utterances, no speech-related tasks were proposed.

This edition saw an increase in tasks related to creativity and style. However, one of such tasks, namely CHANGE-it, had no participation, probably due to its complexity. CHANGE-it required participants to perform some form of style transfer on newspaper articles. This is not a trivial task by itself, and it is arguably rendered even more complex by the little availability of both resources and studies on the subject. Moreover, it is completely novel for the Italian community, but it is expected to grow in interest also thanks to the development and the release of the task dataset. Another task that received a rather low number of submissions due to its complexity is GhigliottinAI. Despite being a rather simple word-correlation problem by itself, it required complex modelling of language and semantics to beat the challenge. A very interesting innovation provided by this task was the evaluation framework, based on APIs, via a Remote Evaluation Server (RES). In general, the most participated tasks have been those by which the linguistic problem could be modelled as a direct classification or regression task. This task also attracted media attention, whereas the workshop event was organized in collaboration with the Italian public broadcast television (RAI), and the task was later mentioned on live TV.

EVALITA 2020 faced unique challenges due to the 2020 COVID-19 pandemic. The workshop was organized as an online event, and the participants were provided a repository to upload video presentations.[6] A virtual space was created on the platform Gather[7] to foster unstructured interaction, nicknamed *Evalitown* (Figure 6). The interaction was lively during the day of the workshop, as was the use of Twitter to promote individual works and tasks.

5 https://huggingface.co/Musixmatch
6 https://vimeo.com/showcase/7882458
7 https://gather.town/

Figure 6
Screenshots of the interaction in *Evalitown*.

The competition attracted a record number of participating teams from academia and industry, for a total of 51 teams and more than 180 authors with affiliations in 18 countries. Hopefully, this means that EVALITA is becoming more and more popular also with foreign contributors, and it is becoming an international workshop. First of all, this success confirms the beneficial impact of the organization of the evaluation period based on non-overlapping windows (adopted from EVALITA 2018) in order to help those who want to participate in more than one task. Moreover, we speculate that the technological advancements and ease of use of existing open-source libraries for machine learning and natural language processing improved the accessibility to the tasks, even for master students. In fact, we noticed an increase in the participation of students, that contributed with state-of-the-art solutions to the tasks. We can argue that the spread of frameworks such as PyTorch and Keras, together with pre-trained, off-the-shelf language models, lowered the set-up costs to deal with complex NLP tasks. In general, we noticed that most of the best systems are based on neural approaches. Among them, BERT or similar Transformer-based architectures achieved the best results: more specifically, at least in 11 out of 14 tasks best results (in at least one sub-task) were obtained by neural architectures based on or combined with Transformers.

We are confident that the positive trends observed in this edition, concerning the participation and the proliferation of tasks, have not yet reached a plateau. It would be desirable, among other aspects, to see more tasks involving challenging settings such as, for example, multi-modal or multi-lingual analysis involving Italian, in future EVALITA editions. Several areas represent fertile ground to organize future tasks, such as domain adaptation (which was considered in previous editions of EVALITA), or few-shot learning to support the definition of robust systems in challenging low-resource settings. Finally, we believe in the importance of defining more structured tasks involving real applications to challenge the Italian community, e.g., Question Answering or Dialogue Agents.

Acknowledgments

The work of Valerio Basile is partially supported by the EVALITA4ELG project (Italian EVALITA Benchmark Linguistic Resources, NLP Services and Tools for the ELG platform), funded by the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 825627 (ELG).

References

Alzetta, Chiara, Alessio Miaschi, Felice Dell'Orletta, Frosina Koceva, and Ilaria Torre. 2020. PRELEARN@EVALITA2020: Overview of the prerequisite relation learning task for italian. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*, volume 2765 of *CEUR Workshop Proceedings*, Online, December. CEUR-WS.org.

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. *Intelligenza Artificiale*, 9(1):43–61.

Attardi, Giuseppe and Maria Simi. 2009. Overview of the evalita 2009 part-of-speech tagging task. In *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*, Reggio Emilia, Italy, December.

Aversano, Guido, Niko Brümmer, and Mauro Falcone. 2009. Evalita 2009 speaker identity verification application track-organizer's report. In *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*, Reggio Emilia, Italy, December.

Badino, Leonardo. 2016. The artiphon task 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 (EVALITA 2016)*, volume 1749 of *CEUR Workshop Proceedings*, Napoli, Italy, December. CEUR-WS.org.

Baggia, Paolo, Francesco Cutugno, Morena Danieli, Roberto Pieraccini, Silvia Quarteroni, Giuseppe Riccardi, and Pierluigi Roberti. 2009. The multi-site 2009 evalita spoken dialog system evaluation. In *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*, Reggio Emilia, Italy, December.

Barbieri, Francesco, 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 (EVALITA 2016)*, volume 1749 of *CEUR Workshop Proceedings*, Napoli, Italy, December. CEUR-WS.org.

Bartalesi Lenzi, Valentina, Manuela Speranza, and Rachele Sprugnoli. 2013. Named entity recognition on transcribed broadcast news at evalita 2011. In Bernardo Magnini, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, editors, *Evaluation of Natural Language and Speech Tools for Italian*, pages 86–97, Berlin, Heidelberg. Springer Berlin Heidelberg.

Bartalesi Lenzi, Valentina and Rachele Sprugnoli. 2007. Description and results of the tern task. *Intelligenza Artificiale*, 4:55–57.

Basile, Pierpaolo, Valerio Basile, Danilo Croce, and Marco Polignano. 2018. 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 Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2263 of *CEUR Workshop Proceedings*, Turin, Italy, December. CEUR-WS.org.

Basile, Pierpaolo, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020. DIACR-Ita@EVALITA2020: Overview of the evalita2020 diachronic lexical semantics (diacr-ita) task. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*, volume 2765 of *CEUR Workshop Proceedings*, Online, December. CEUR-WS.org.

Basile, Pierpaolo, 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 (EVALITA 2016)*, volume 1749 of *CEUR Workshop Proceedings*, Napoli, Italy, December. CEUR-WS.org.

Basile, Pierpaolo, Danilo Croce, Valerio Basile, 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 Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2263 of *CEUR Workshop Proceedings*, Turin, Italy, December. CEUR-WS.org.

Basile, Pierpaolo, 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 Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2263 of *CEUR Workshop Proceedings*, Turin, Italy, December. CEUR-WS.org.

Basile, Pierpaolo, Marco Lovetere, Johanna Monti, Antonioa Pascucci, Federico Sangati, and Lucia Siciliani. 2020. Ghigliottin-AI@EVALITA2020: Evaluating artificial players for the language game "la ghigliottina". In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*, volume 2765 of *CEUR Workshop Proceedings*, Online, December. CEUR-WS.org.

Basile, Pierpaolo 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 Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2263 of *CEUR Workshop Proceedings*, Turin, Italy, December. CEUR-WS.org.

Basile, Pierpaolo, Viviana Patti, Francesco Cutugno, Malvina Nissim, and Rachele Sprugnoli. 2017. Evalita goes social: Tasks, data. *IJCoL. Italian Journal of Computational Linguistics*, 3(3-1):93–127.

Basile, Valerio, Andrea Bolioli, Viviana Patti, Paolo Rosso, and Malvina Nissim. 2014. Overview of the evalita 2014 sentiment polarity classification task. In Cristina Bosco, Piero Cosi, Felice Dell'Orletta, Mauro Falcone, Simonetta Montemagni, and Maria Simi, editors, *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & the Fourth International Workshop EVALITA 2014*, volume 2, pages 50–57, Pisa, Italy, December. Pisa University Press.

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 Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online, December. CEUR.org.

Basili, Roberto, Diego De Cao, Alessandro Lenci, Alessandro Moschitti, and Giulia Venturi. 2013. Evalita 2011: The frame labelingover italian texts task. In Bernardo Magnini, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, editors, *Evaluation of Natural Language and Speech Tools for Italian*, pages 195–204, Berlin, Heidelberg. Springer Berlin Heidelberg.

Bentivogli, Luisa, Alessandro Marchetti, and Emanuele Pianta. 2013. The news people search task at evalita 2011: Evaluating cross-document coreference resolution of named person entities in italian news. In Bernardo Magnini, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, editors, *Evaluation of Natural Language and Speech Tools for Italian*, pages 126–134, Berlin, Heidelberg. Springer Berlin Heidelberg.

Bertagna, Francesca, Antonio Toral, and Nicoletta Calzolari. 2007. The all-words wsd task. *Intelligenza Artificiale*, 4:50–52.

Bos, Johan, Fabio Massimo Zanzotto, and Marco Pennacchiotti. 2009. Textual entailment at evalita 2009. In *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*, Reggio Emilia, Italy, December.

Bosco, Cristina, Silvia Ballarè, Massimo Cerruti, Eugenio Goria, and Caterina Mauri. 2020. KIPoS@EVALITA2020:overview of the task on kiparla part of speech tagging. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*, volume 2765 of *CEUR Workshop Proceedings*, Online, December. CEUR-WS.org.

Bosco, Cristina, Felice Dell'Orletta, Simonetta Montemagni, Manuela Sanguinetti, and Maria Simi. 2014. The evalita 2014 dependency parsing task. In Cristina Bosco, Piero Cosi, Felice Dell'Orletta, Mauro Falcone, Simonetta Montemagni, and Maria Simi, editors, *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & the Fourth International Workshop EVALITA 2014*, volume 2, pages 1–8, Pisa, Italy, December. Pisa University Press.

Bosco, Cristina, Tamburini Fabio, Bolioli Andrea, 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 (EVALITA 2016)*, volume 1749 of *CEUR Workshop Proceedings*, Napoli, Italy, December. CEUR-WS.org.

Bosco, Cristina, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. 2018. Overview of the evalita 2018 hate speech detection task. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2263 of *CEUR Workshop Proceedings*, Turin, Italy, December. CEUR-WS.org.

Bosco, Cristina and Alessandro Mazzei. 2013. The evalita dependency parsing task: From 2007 to 2011. In Bernardo Magnini, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, editors, *Evaluation of Natural Language and Speech Tools for Italian*, pages 1–12, Berlin, Heidelberg. Springer Berlin Heidelberg.

Bosco, Cristina, Alessandro Mazzei, and Alberto Lavelli. 2013. Looking back to the evalita constituency parsing task: 2007-2011. In Bernardo Magnini, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, editors, *Evaluation of Natural Language and Speech Tools for Italian*, pages 46–57, Berlin, Heidelberg. Springer Berlin Heidelberg.

Bosco, Cristina, Alessandro Mazzei, and Vincenzo Lombardo. 2007. An analysis of the first parsing system contest for italian. *Intelligenza Artificiale*, 4:30–33.

Bosco, Cristina, Alessandro Mazzei, and Vincenzo Lombardo. 2009. Evalita'09 parsing task: constituency parsers and the penn format for italian. In *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*, Reggio Emilia, Italy, December.

Bosco, Cristina, Simonetta Montemagni, Alessandro Mazzei, Vincenzo Lombardo, Felice Dell'Orletta, and Alessandro Lenci. 2009. Evalita'09 parsing task: comparing dependency parsers and treebanks. In *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*, Reggio Emilia, Italy, December.

Brunato, Dominique, Cristiano Chesi, Felice Dell'Orletta, Simonetta Montemagni, Giulia Venturi, and Roberto Zamparelli. 2020. AcCompl-it@EVALITA2020: Overview of the acceptability & complexity evaluation task for italian. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online, December. CEUR.org.

Brutti, Alessio, Mirco Ravanelli, and Maurizio Omologo. 2014. Saslodom: Speech activity detection and speaker localization in domestic environments. In Cristina Bosco, Piero Cosi, Felice Dell'Orletta, Mauro Falcone, Simonetta Montemagni, and Maria Simi, editors, *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & the Fourth International Workshop EVALITA 2014*, volume 2, pages 139–146, Pisa, Italy, December. Pisa University Press.

Caputo, Annalina, Marco de Gemmis, Pasquale Lops, Francesco Lovecchio, Vito Manzari, and Acquedotto Pugliese AQP Spa. 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 (EVALITA 2016)*, volume 1749 of *CEUR Workshop Proceedings*, Napoli, Italy, December. CEUR-WS.org.

Caselli, Tommaso, Nicole Novielli, Viviana Patti, and Paolo Rosso. 2018. EVALITA 2018: Overview of the 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2263 of *CEUR Workshop Proceedings*, Turin, Italy, December. CEUR-WS.org.

Cignarella, Alessandra Teresa, Simona Frenda, Valerio Basile, Cristina Bosco, Viviana Patti, Paolo Rosso, et al. 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 Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), volume 2263 of *CEUR Workshop Proceedings*, Turin, Italy, December. CEUR-WS.org.

Cignarella, Alessandra Teresa, Mirko Lai, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2020. SardiStance@EVALITA2020: Overview of the task on stance detection in italian tweets. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*, volume 2765 of *CEUR Workshop Proceedings*, Online, December. CEUR-WS.org.

Cimino, Andrea, Felice Dell'Orletta, and Malvina Nissim. 2020. TAG-it@EVALITA2020: Overview of the topic, age, and gender prediction task for italian. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*, volume 2765 of *CEUR Workshop Proceedings*, Online, December. CEUR-WS.org.

Coro, Gianpaolo, Roberto Gretter, and Marco Matassoni. 2009. Evalita 2009: Description and results of the speech recognition task. In *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*, Reggio Emilia, Italy, December.

Cosi, Piero, Francesco Cutugno, Vincenzo Galatà, and Antonio Origlia. 2014. Forced alignment on children speech. In Cristina Bosco, Piero Cosi, Felice Dell'Orletta, Mauro Falcone, Simonetta Montemagni, and Maria Simi, editors, *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & the Fourth International Workshop EVALITA 2014*, volume 2, pages 124–126, Pisa, Italy, December. Pisa University Press.

Cutugno, Francesco, Maria Di Maro, Sara Falcone, Marco Guerini, Bernardo Magnini, Antonio Origlia, Fondazione Bruno Kessler, and Trento—Italy Povo. 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 Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2263 of *CEUR Workshop Proceedings*, Turin, Italy, December. CEUR-WS.org.

Cutugno, Francesco, Antonio Origlia, and Dino Seppi. 2013. Evalita 2011: Forced alignment task. In Bernardo Magnini, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, editors, *Evaluation of Natural Language and Speech Tools for Italian*, pages 305–311, Berlin, Heidelberg. Springer Berlin Heidelberg.

De Mattei, Lorenzo, Michele Cafagna, Felice Dell'Orletta, Malvina Nissim, and Albert Gatt. 2020a. CHANGE-IT@EVALITA2020: Change headlines, adapt news, generate. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*, volume 2765 of *CEUR Workshop Proceedings*, Online, December. CEUR-WS.org.

De Mattei, Lorenzo, Graziella De Martino, Andrea Iovine, Alessio Miaschi, Marco Polignano, and Giulia Rambelli. 2020b. ATE_ABSITA@EVALITA2020: Overview of the aspect term extraction and aspect-based sentiment analysis task. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*, volume 2765 of *CEUR Workshop Proceedings*, Online, December. CEUR-WS.org.

Dei Rossi, Stefano, Giulia Di Pietro, and Maria Simi. 2013. Description and results of the supersense tagging task. In Bernardo Magnini, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, editors, *Evaluation of Natural Language and Speech Tools for Italian*, pages 166–175, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dell'Orletta, Felice, Simone Marchi, Simonetta Montemagni, Giulia Venturi, Tommaso Agnoloni, and Enrico Francesconi. 2013. Domain adaptation for dependency parsing at evalita 2011. In Bernardo Magnini, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, editors, *Evaluation of Natural Language and Speech Tools for Italian*, pages 58–69, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dell'Orletta, Felice 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 Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2263 of *CEUR Workshop Proceedings*, Turin, Italy, December. CEUR-WS.org.

Dell'Orletta, Felice 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 Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2263 of *CEUR Workshop Proceedings*, Turin, Italy, December. CEUR-WS.org.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Di Maro, Maria, II di Napoli'Federico, 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 Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2263 of *CEUR Workshop Proceedings*, Turin, Italy, December. CEUR-WS.org.

Fersini, Elisabetta, Debora Nozza, and Paolo Rosso. 2018. 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 Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2263 of *CEUR Workshop Proceedings*, Turin, Italy, December. CEUR-WS.org.

Fersini, Elisabetta, Debora Nozza, and Paolo Rosso. 2020. AMI@EVALITA2020: Automatic misogyny identification. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*, volume 2765 of *CEUR Workshop Proceedings*, Online, December. CEUR-WS.org.

Giorgioni, Simone, Marcello Politi, Samir Salman, Danilo Croce, and Roberto Basili. 2020. Unitor@ sardistance2020: Combining transformerbased architectures and transfer learning for robust stance detection. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*, volume 2765 of *CEUR Workshop Proceedings*, Online, December. CEUR-WS.org.

Gregori, Lorenzo, Maria Montefinese, Daniele P. Radicioni, Andrea Amelio Ravelli, and Rossella Varvara. 2020. CONCRETEXT@EVALITA2020: The concreteness in context task. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*, volume 2765 of *CEUR Workshop Proceedings*, Online, December. CEUR-WS.org.

Lenzi, Valentina Bartalesi and Rachele Sprugnoli. 2009. Evalita 2009: Description and results of the local entity detection and recognition (ledr) task. In *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*, Reggio Emilia, Italy, December.

Massidda, Riccardo. 2020. rmassidda@ dadoeval: Document dating using sentence embeddings at evalita 2020. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*, volume 2765 of *CEUR Workshop Proceedings*, Online, December. CEUR-WS.org.

Matassoni, Marco, Fabio Brugnara, and Roberto Gretter. 2013. Evalita 2011: Automatic speech recognition large vocabulary transcription. In Bernardo Magnini, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, editors, *Evaluation of Natural Language and Speech Tools for Italian*, pages 274–285, Berlin, Heidelberg. Springer Berlin Heidelberg.

Menini, Stefano, Giovanni Moretti, Rachele Sprugnoli, and Sara Tonelli. 2020. DaDoEval@EVALITA2020: Same-genre and cross-genre dating of historical documents. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*, volume 2765 of *CEUR Workshop Proceedings*, Online, December. CEUR-WS.org.

Mikolov, Tomás, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013*, Scottsdale, Arizona, USA, May.

Mikolov, Tomás, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings*, pages 3111–3119, Lake Tahoe, Nevada, United States, December 5-8, 2013.

Miliani, Martina, Giulia Giorgi, Ilir Rama, Guido Anselmi, and Gianluca E. Lebani. 2020. DANKMEMES@EVALITA2020: The memeing of life: memes, multimodality and politics. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*, volume 2765 of *CEUR Workshop Proceedings*, Online, December. CEUR-WS.org.

Minard, Anne-Lyse, Manuela Speranza, Tommaso Caselli, and Fondazione Bruno Kessler. 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 (EVALITA 2016)*, volume 1749 of *CEUR Workshop Proceedings*, Napoli, Italy, December. CEUR-WS.org.

Origlia, Antonio and Vincenzo Galata. 2014. Evalita 2014: Emotion recognition task (ert). In Cristina Bosco, Piero Cosi, Felice Dell'Orletta, Mauro Falcone, Simonetta Montemagni, and Maria Simi, editors, *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & the Fourth International Workshop EVALITA 2014*, volume 2, Pisa, Italy, December. Pisa University Press.

Romano, Antonio and Claudio Russo. 2014. Human and machine language/dialect identification from natural speech and artificial stimuli: a pilot study with italian listener. In Cristina Bosco, Piero Cosi, Felice Dell'Orletta, Mauro Falcone, Simonetta Montemagni, and Maria Simi, editors, *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & the Fourth International Workshop EVALITA 2014*, volume 2, pages 131–138, Pisa, Italy, December. Pisa University Press.

Romito, Luciano and Vincenzo Galatà. 2009. Forensic speaker identity verification (f-siv) in italy first evaluation campaign evalita-2009. In *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*, Reggio Emilia, Italy, December.

Ronzano, Francesco, Francesco Barbieri, Endang Wahyu Pamungkas, Viviana Patti, Francesca Chiusaroli, et al. 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 Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2263 of *CEUR Workshop Proceedings*, Turin, Italy, December. CEUR-WS.org.

Sanguinetti, Manuela, Gloria Comandini, Elisa Di Nuovo, Simona Frenda, Marco Stranisci, Cristina Bosco, Tommaso Caselli, Viviana Patti, and Irene Russo. 2020. HaSpeeDe 2@EVALITA2020: Overview of the evalita 2020 hate speech detection task. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*, volume 2765 of *CEUR Workshop Proceedings*, Online, December. CEUR-WS.org.

Sarti, Gabriele. 2020. Umberto-mtsa@ accompl-it: Improving complexity and acceptability prediction with multi-task learning on self-supervised annotations. *arXiv preprint arXiv:2011.05197*.

Speranza, Manuela. 2007. The named entity recognition task. *Intelligenza Artificiale*, 4:66–68.

Speranza, Manuela. 2009. The named entity recognition task at evalita 2009. In *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*, Reggio Emilia, Italy, December.

Tamburini, Fabio. 2007. The part-of-speech tagging task. *Intelligenza Artificiale*, 4:4–7.

Tamburini, Fabio. 2013. The lemmatisation task at the evalita 2011 evaluation campaign. In Bernardo Magnini, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, editors, *Evaluation of Natural Language and Speech Tools for Italian*, pages 230–238, Berlin, Heidelberg. Springer Berlin Heidelberg.

Tommaso, Caselli, R Sprugnoli, Speranza Manuela, and Monachini Monica. 2014. Eventi evaluation of events and temporal information at evalita 2014. In Cristina Bosco, Piero Cosi,

Felice Dell'Orletta, Mauro Falcone, Simonetta Montemagni, and Maria Simi, editors, *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & the Fourth International Workshop EVALITA 2014*, volume 2, pages 27–34, Pisa, Italy, December. Pisa University Press.

Toral, Antonio. 2009. The lexical substitution task at evalita 2009. In *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*, Reggio Emilia, Italy, December.

Uryupina, Olga and Massimo Poesio. 2013. Evalita 2011: Anaphora resolution task. In Bernardo Magnini, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, editors, *Evaluation of Natural Language and Speech Tools for Italian*, pages 146–155, Berlin, Heidelberg. Springer Berlin Heidelberg.

Appendix A: EVALITA Edition 2007-2020

Edition	Area	Tasks
2007	NER, Events and Time	Temporal Expression Recognition and Normalization (Bartalesi Lenzi and Sprugnoli 2007), Named Entity Recognition (Speranza 2007)
	Parsing	Parsing (Bosco, Mazzei, and Lombardo 2007)
	PoS Tagging	Part of Speech Tagging (Tamburini 2007)
	Senses and Frames	Word Sense Disambiguation (Bertagna, Toral, and Calzolari 2007)
2009	NER, Events and Time	Entity Recognition (Speranza 2009; Lenzi and Sprugnoli 2009)
	Parsing	Parsing (Bosco et al. 2009; Bosco, Mazzei, and Lombardo 2009)
	PoS Tagging	PoS-Tagging (Attardi and Simi 2009)
	Senses and Frames	Lexical Substitution (Toral 2009)
	Speech	Connected Digits Recognition (Coro, Gretter, and Matassoni 2009),Spoken Dialogue Systems Evaluation (Baggia et al. 2009),Speaker Identity Verification (Aversano, Brümmer, and Falcone 2009; Romito and Galatà 2009)
	Textual Entailment	Textual Entailment (Bos, Zanzotto, and Pennacchiotti 2009)
2011	Coreference	Cross-document Coreference Resolution (Bentivogli, Marchetti, and Pianta 2013), Anaphora Resolution (Uryupina and Poesio 2013)
	Lemmatisation	Lemmatisation (Tamburini 2013)
	NER, Events and Time	Named Entity Recognition on Transcribed Broadcast News (Bartalesi Lenzi, Speranza, and Sprugnoli 2013)
	Parsing	Parsing (Bosco and Mazzei 2013; Bosco, Mazzei, and Lavelli 2013),Domain Adaptation for Dependency Parsing (Dell'Orletta et al. 2013)
	Senses and Frames	Super Sense Tagging (Dei Rossi, Di Pietro, and Simi 2013), Frame Labeling over Italian Texts (Basili et al. 2013)
	Speech	Automatic Speech Recognition - Large Vocabulary Transcription (Matassoni, Brugnara, and Gretter 2013), Forced Alignment on Spontaneous Speech (Cutugno, Origlia, and Seppi 2013),Voice Applications on Mobile
2014	Affect	Sentiment Polarity Classification (Basile et al. 2014), Emotion Recogniton Task (Origlia and Galata 2014)
	NER, Events and Time	Evaluation of Events and Temporal Information (Tommaso et al. 2014)
	Parsing	Dependency Parsing (Bosco et al. 2014)
	Speech	Forced Alignment on Children Speech (Cosi et al. 2014), Human and Machine Dialect Identification from Natural Speech and Artificial Stimuli (Romano and Russo 2014), Speech Activity Detection and Speaker Localization in Domestic Environments (Brutti, Ravanelli, and Omologo 2014)
2016	Affect	SENTIPOLC - SENTIment POLarity Classification (Barbieri et al. 2016)
	Dialog	QA4FAQ - Question Answering for Frequently Asked Questions (Caputo et al. 2016)
	NER, Events and Time	FactA - Event Factuality Annotation (Minard et al. 2016), NEEL-IT - Named Entity rEcognition and Linking in Italian Tweets (Basile et al. 2016)
	PoS Tagging	PoSTWITA - POS tagging for Italian Social Media Texts (Bosco et al. 2016)
	Speech	ArtiPhon - Articulatory Phone Recognition (Badino 2016)

Edition	Area	Tasks
2018	Affect	ABSITA - Aspect-based Sentiment Analysis (Basile et al. 2018a), ITA-Moji - Italian Emoji Prediction (Ronzano et al. 2018), IronITA - Irony Detection in Twitter (Cignarella et al. 2018), AMI - Automatic Misogyny Identification (Fersini, Nozza, and Rosso 2018), HaSpeeDe - Hate Speech Detection (Bosco et al. 2018)
	Dialog	iLISTEN - itaLIan Speech acT labEliNg (Basile and Novielli 2018), IDIAL - Italian DIALogue systems evaluation (Cutugno et al. 2018)
	Semantics	NLP4FUN - Solving language games (Basile et al. 2018b)
	Speech	SUGAR - Spoken Utterances Guiding chef's Assistant Robots (Di Maro et al. 2018)
	Style	GxG - Cross-Genre Gender Prediction (Dell'Orletta and Nissim 2018)
2020	Affect	ATE_ABSITA - Aspect Term Extraction and Aspect-Based Sentiment Analysis (De Mattei et al. 2020b), AMI - Automatic Misogyny Identification (Fersini, Nozza, and Rosso 2020), SardiStance - Stance Detection (Cignarella et al. 2020), HaSpeeDe - Hate Speech Detection (Sanguinetti et al. 2020)
	Multimodality	DANKMEMES - Multimodal Artefacts Recognition (Miliani et al. 2020)
	PoS Tagging	KIPoS - Part-of-speech Tagging on Spoken Language (Bosco et al. 2020)
	Semantics	CONcreTEXT - Concreteness in Context (Gregori et al. 2020), Ghigliottin-AI - Evaluating Artificial Players for the Language Game "La Ghigliottina" (Basile et al. 2020), PRELEARN - Prerequisite Relation Learning (Alzetta et al. 2020)
	Style	CHANGE-IT - Style Transfer (De Mattei et al. 2020a),TAG-it - Topic, Age and Gender Prediction (Cimino, Dell'Orletta, and Nissim 2020), AcCompl-it- Acceptability & Complexity evaluation (Brunato et al. 2020)
	Time and Diacrony	DaDoEval - Dating Documents (Menini et al. 2020), DIACR-Ita - Diachronic Lexical Semantics (Basile et al. 2020)