State of the Art Language Technologies for Italian: The EVALITA 2014 Perspective

Intelligenza Artificiale 9 (2015) 43–61
DOI 10.3233/IA-150076
IOS Press

State of the Art Language Technologies for Italian: The EVALITA 2014 Perspective

Giuseppe Attardi[a], Valerio Basile[b], Cristina Bosco[c], Tommaso Caselli[d], Felice Dell'Orletta[e], Simonetta Montemagni[e,*], Viviana Patti[c], Maria Simi[a] and Rachele Sprugnoli[f]
[a]*Dipartimento di Informatica, Università di Pisa, largo B. Pontecorvo, Pisa, Italy*
[b]*Center for Language and Cognition Groningen, University of Groningen, Broerstraat, Groningen, The Netherlands*
[c]*Dipartimento di Informatica, Università di Torino, corso Svizzera, Torino, Italy*
[d]*Computational Lexicology & Terminology Lab (CLTL), VU Amsterdam, De Boelelaan, HV, Amsterdam, The Netherlands*
[e]*Istituto di Linguistica Computazionale "Antonio Zampolli" - CNR, Via G. Moruzzi, Pisa, Italy*
[f]*FBK - University of Trento, Via Sommariva, Trento, Italy*

Abstract. Shared task evaluation campaigns represent a well established form of competitive evaluation, an important opportunity to propose and tackle new challenges for a specific research area and a way to foster the development of benchmarks, tools and resources. The advantages of this approach are evident in any experimental field, including the area of Natural Language Processing. An outlook on state–of–the–art language technologies for Italian can be obtained by reflecting on the results of the recently held workshop "Evaluation of NLP and Speech Tools for Italian", EVALITA 2014. The motivations underlying individual shared tasks, the level of knowledge and development achieved within each of them, the impact on applications, society and economy at large as well as directions for future research will be discussed from this perspective.

Keywords: Evaluation Campaign, Natural Language Processing, Dependency Parsing, Sentiment Analysis, Temporal Processing

1. Introduction

Evaluation of achieved results is a crucial process of scientific research. This also applies to the area of Natural Language Processing (NLP): establishing a well–grounded evaluation methodology makes it easier to track advances in the field and to assess the impact of the work done. In addition, it can set evaluation standards that can also be exported to other fields. The comparison of the results of different systems is not a trivial task as many parameters can affect and influence this process.

To overcome this issue, over the last ten years shared task evaluation campaigns started being increasingly popular as a competitive form of evaluation. Shared task evaluation campaigns represent an important opportunity to investigate ways to tackle the challenges a specific research area is facing, where different approaches to a well–defined problem are compared based on their performance on the same task with respect to the same dataset. The datasets used within evaluation campaigns become reference resources of the scientific community and are used to assess effectiveness and performance of a given system or technology with respect to a specific task.

International evaluation campaigns held since the 90s, starting from the pioneer initiatives represented by the Automatic Content Extraction (ACE) evaluation campaign[1] and the Message Understanding Conferences (MUC)[2] to the more recent CoNLL shared Tasks[3]

*Corresponding author: Simonetta Montemagni, Istituto di Linguistica Computazionale "Antonio Zampolli" – CNR, Via G. Moruzzi 1, I-56124, Pisa, Italy. Tel.: +39 050 3152850; Fax: +39 050 3152839; E-mail: simonetta.montemagni@ilc.cnr.it.

[1] http://www.itl.nist.gov/iad/mig/tests/ace/

[2] http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html

[3] http://ifarm.nl/signll/conll/

and the SemEval workshops[4], are considered as a point of reference for Natural Language Processing research because of their valuable contribution in defining and improving state–of–the–art technologies. In particular, their impact on the research carried out by the scientific community can be summarized as follows. On the one hand, the comparison of the performance of systems against shared datasets and benchmarks improves the significance of the results obtained by participants and establishes acknowledged baselines. On the other hand, the development of the datasets needed for training and/or testing the participant systems is in itself an activity which positively influences the area, widening the availability of reliable resources to be exploited for both evaluation and application purposes.

Since 2007, EVALITA[5], a periodic evaluation campaign of Natural Language Processing and Speech Technologies for the Italian language, has been organized. Each edition of the campaign, held in 2007 [60], 2009 [61], 2011 [45] and 2014 [21], has been organized around a set of shared tasks dealing with both spoken and written language[6] and varying with respect to the challenges tackled and the datasets used. In each edition, the selection of tasks has been carried out by taking into account different factors, among which a crucial role is played by the possible impact of the proposed task on society and economy. Previous editions of the EVALITA campaign have focused mainly on standard tasks in NLP such as Parsing (costituency and dependency parsing), Word Sense Disambiguation, Part-of-Speech (PoS) Tagging, Named Entity Recognition and Semantic Role Labelling.

In this paper, we will focus on tasks concerned with the processing of written language[7]. In the EVALITA 2014 edition three tasks had been selected for this area, namely:

1. dependency parsing (see Section 2), a well–known task in NLP which has been organized around a newly developed resource based on the Stanford Dependencies annotation scheme [31], which is gaining popularity as a formalism suitable for further semantic processing and information extraction;
2. temporal processing (see Section 3), a task which aims at extracting temporal information (i.e. events, temporal expressions and temporal relations) from a document and which can be used for practical activities in lots of domains such as monitoring clinical data, tracking opinions in time, and developing support decision systems;
3. the classification of polarity of social media posts (see Section 4), an activity that is known as crucial for social media monitoring platforms providing business services, monitoring political sentiment, extracting critical information during times of mass emergency, or detecting moods and happiness in social media services.

Results achieved within each single task range from the advancement of state–of–the–art technologies for Italian to the development of the linguistic resources needed to carry out the task. Last but not least, it is worth emphasizing here that the relevance of an evaluation campaign goes beyond the borders of the community working on a specific language. On the one hand, models and tools developed for other languages can be applied to the language on which the evaluation campaign focuses, thus paving the way to a reliable and sound cross–linguistic comparison. On the other hand, technologies originally developed for Italian can be extended to deal with other languages.

2. Dependency Parsing – DP

Dependency parsing aims at reconstructing the syntactic structure of a sentence, represented in terms of binary asymmetric relations, called *dependencies*, that link a word (the *head*) to a syntactically subordinate word (the *dependent*). The set of these relations forms a tree that can be pictured like in Fig. 1.

The figure shows the dependency parse tree for the sentence *Del legno mette in risalto le venature*

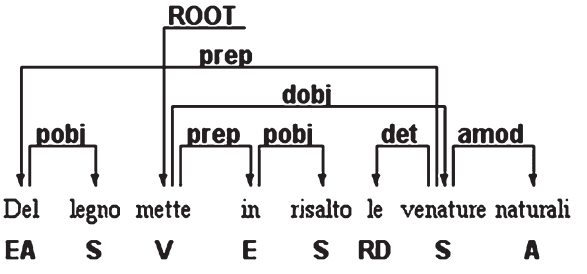

Fig. 1. Example of dependency tree.

[4] http://www.aclweb.org/aclwiki/index.php?title=SemEval_Portal

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

[6] Speech tasks started being organized starting from the second edition, held in 2009.

[7] For tasks focusing on speech technologies, the interested reader is referred to the speech tasks section of [21].

naturali[8]. The arcs in the tree are labeled with tags that denote the dependency types: for example, the word *venature* is tagged as the direct object (`dobj`) of the verb *mette*.

An advantage of the dependency formalism is that in principle all that is needed to produce a parse tree is to decide whether a word is dependent on another, i.e. parsing can be reduced to a set of binary decisions for each pair of words in a sentence. In practice lots of research has been devoted to streamline the process and to exploit contextual information for improving the decision accuracy. Statistical parsers can be trained to learn how to perform these decisions from a corpus of sentences annotated with dependencies.

In recent years, dependency corpora have become more readily available for many languages. A single parser can be trained to analyse multiple languages, with minor adaptation efforts.

The task of dependency parsing has been investigated for multiple languages within the evaluation campaign of NLP systems at the CoNLL-X conference [22], and later editions [36, 75]. Shared tasks focusing on dependency parsing have been organized e.g. for individual languages (e.g. Indian languages [40] or German [44]), non-canonical variants of languages [54] as well as for morphologically rich languages [70]. EVALITA has organized dependency parsing tasks focusing on Italian since 2007 [16]. These campaigns fostered both the creation of dependency treebanks and the development of dependency parsers exploiting different approaches. The results of the evaluations show a constant increase in accuracy and coverage. However, dependency parsing still represents an active research area since there are a number of open issues and potential for improvements. For example:

- integrating parsing with other layers of analysis, e.g. PoS tagging and Semantic Role Labelling;
- developing new strategies for applying transitions and improving the accuracy;
- exploiting distributed word representations (word embeddings) as features.

Moreover, attention is also shifting from the analysis of just grammatical structure to sentence semantics, as addressed for example in the SemEval 2014 task "Broad-Coverage Semantic Dependency Parsing" [52], that aims at identifying semantic dependencies for all content words in a sentence.

[8] Lit. *Of the wood (it) enhances the veneers natural* meaning that *(it) enhances the natural veneers of the wood.*

In order to test the suitability of dependency structures for inferring meaning representations, EVALITA 2014 introduced a new evaluation metric for assessing the suitability of dependency parse trees for information extraction: this can be seen as a first step towards producing semantically–oriented representations.

Another promising line of research in the area of dependency parsing is represented by "cross–lingual dependency parsing", where annotated resources from one language are used to learn models for another: in this way, the range and variety of languages for which dependency parsers can be developed is significantly extended also to cover under–resourced ones.

EVALITA 2014 has hence introduced a dependency parsing task targeted at experiments for carrying linguistic knowledge about dependency structures across languages. In particular, the Italian corpus was used as a source of dependency structures to be identified in texts encoded in other languages.

2.1. Dataset

Stanford Dependencies (SD) have gained the status of *de facto* standard for dependency–based treebank annotations [30]. The *Italian Stanford Dependency Treebank* (ISDT) [18, 71] resulted from the harmonization and merging of smaller existing Italian treebanks (namely, TUT and ISST–TANL, already used in previous EVALITA campaigns) originally adopting incompatible annotation schemes [17]. ISDT includes texts representative of various textual genres, ranging from legal texts (the Civil code, the Italian Constitution and European directives) to newspaper and Wikipedia articles, for a total of around 171,500 word tokens.

The annotation scheme used for ISDT follows as closely as possible the specifications provided in the SD manual for English, with few variations accounting for syntactic peculiarities of the Italian language: the Italian localization of the Stanford Dependency scheme is described in detail in [18]. The tagset used, consisting in 41 dependency tags, and the Italian-specific annotation guidelines are reported in a dedicated webpage[9]. It is also worth pointing out here that the Stanford Dependency scheme has been defined at the level of both dependency tagset and head selection criteria with a specific view to supporting Information Extraction tasks: this makes ISDT the most suitable resource to be used for the dependency parsing shared task organized in EVALITA 2014.

[9] See:http://medialab.di.unipi.it/wiki/ISDT

In EVALITA 2014 different variants of the ISDT resource were exploited, obtained through an automatic conversion process. From the basic ISDT variant, two different versions of the resource were generated: a semantically–oriented representation (so–called *collapsed and propagated* in SD parlance [31]), and a universal version (henceforth referred to as "uISDT") conforming to the Universal Stanford Dependencies scheme defined in the framework of *The Universal Dependency Treebank Project*[10].

2.2. Description of DP tasks

The dependency parsing task of EVALITA 2014 was split into two main subtasks: the first, henceforth referred to as *Dependency Parsing for Information Extraction* or DPIE, is a basic subtask focusing on standard dependency parsing of Italian texts, with a dual evaluation track aimed at testing both the performance of parsing systems and their suitability to Information Extraction tasks; the second subtask, i.e. *Cross–Language dependency Parsing* or CLaP, is a pilot multilingual task where a source Italian treebank is used to train a parsing model which is then used to parse other (not necessarily typologically related) languages. This twofold organization was meant to advance research in currently hot areas of dependency parsing by exploiting the new ISDT resource and a set of multilingual treebanks from *The Universal Dependency Treebank Project* [48].

2.2.1. The DPIE task

DPIE was organized as a classical dependency parsing task, where the performance of different parsers is compared on the basis of the same set of test data provided by the organizers. In order to allow participants to develop and tune their systems, the ISDT resource was split into a training set (165,975 tokens) and a validation set (12,578 tokens). For the purposes of the final evaluation, a new test data set was developed, for a total of 9,442 tokens covering different textual genres.

The main novelty of this subtask consisted in the methodology adopted for evaluating the outputs of participant systems. In addition to the Labeled Attachment Score (LAS) and Unlabelled Attachment Score (UAS), which represent standard metrics in dependency parsing, an alternative and semantically–oriented metric was introduced to assess the ability of participant systems to produce suitable and accurate output for information extraction applications. The semantically–oriented evaluation was carried out against the *collapsed and propagated* version of the parsers output and was based on a subset (19 out of 41) of the relation types selected as more relevant, i.e. semantically–loaded: this is the case of relations typically linking content words, e.g. classical dependency relations such as subject (`nsubj`), direct object (`dobj`) or adjectival complement (`acomp`) as well as semantic relations such as temporal modifier (`tmod`). In this case, used evaluation metrics were *Precision*, *Recall* and *F1* measures.

2.2.2. The CLaP task

CLaP is a cross-lingual transfer parsing task, organized along the lines of the experiments described in McDonald et al. [48]. In this task, participants were asked to use their parsers trained on uISDT on test sets of other languages, annotated according to the Universal Dependency Treebank Project guidelines.

The languages involved in the task are all the languages distributed from the Universal Dependency Treebank Project with the exclusion of Italian, i.e.: Brazilian-Portuguese, English, Finnish, French, German, Indonesian, Japanese, Korean, Spanish and Swedish. For this task, participant systems were provided with:

- a development set consisting of uISDT, the universal version of ISDT obtained through automatic conversion, and validation sets of about 7,500 tokens for each of the remaining ten languages of the Universal Dependency Treebank;
- a number of test sets (one for each target language) of about 7,500 tokens for the evaluation, with gold PoS and morphology and without dependency information. Test sets were built by randomly extracting sentences from SD treebanks available at https://code.google.com/p/uni-dep-tb/.

The use of external resources (e.g. dictionaries, lexicons, machine translation outputs, etc.) in addition to the corpus provided for training was allowed. Participants in this task were also allowed to focus on a subset of languages only.

2.3. Participant systems and results

Four teams submitted runs to the DPIE task: Attardi et al. (University of Pisa), Lavelli (FBK, Trento), Mazzei (University of Torino) and Grella (Parsit, Torino). A detailed description of their experiments and

[10] https://code.google.com/p/uni-dep-tb/

Table 1
DPIE: LAS and UAS scores of the submitted runs

Submission	LAS	UAS
Attardi run1	**87.89**	**90.16**
Attardi run3	87.84	90.15
Attardi run2	87.83	90.06
Lavelli run3	87.53	89.90
Lavelli run2	87.37	89.94
Mazzei run1	87.21	89.29
Mazzei run2	87.05	89.48
Lavelli run1	86.79	89.14
Grella	84.72	90.03

Table 2
Results on the English Penn TreeBank converted to Stanford Dependencies. Starred values are for experiments using external resources

Parser	LAS	UAS
MaltParser	86.46	89.20
MSTParser	87.53	90.87
DeSR	89.38	91.18
[28]	89.60*	91.80*
TurboParser	89.67	92.20

approaches is presented in the proceedings of EVALITA 2014 [21].

Participants used various publicly available state-of-the-art parsers, namely DeSR parser [4], MALT parser [51], MATE parser [15], TurboParser [46] and ZPar [81]. Several runs where obtained by ensemble combinations of multiple parser outputs. Table 1 summarizes the accuracy scores of all submitted runs measured in terms of LAS/UAS scores (best scores are highlighted in bold). The top four results are grouped together, since their score difference is not statistically significant ($p < 0.05$).

To compare these results with the state of the art dependency parsing in other languages, we report in Table 2 the accuracy of a few parsers on the English Penn Treebank converted to Basic Stanford Dependencies using Stanford rules: for details see [43]. Since these results were not obtained within the controlled settings of an evaluation task, they should be taken as indicative.

Slightly higher accuracies for English are to be expected for several reasons: the English treebank is about 7.5 times bigger than the current EVALITA corpus and English is grammatically simpler, with stricter word order and less rich morphology.

The result by [28] uses slightly different settings, but we report it here since it exploits *word embeddings* as features in a transition based parser. The effectiveness of embeddings varies with different languages: experiments by one of the authors showed a significant drop in accuracy using that parser with Italian embeddings created from the Italian Wikipedia. The approach though is promising and deserves further investigation.

Table 3
DPIE subtask: scores of all submissions on the selected relations, computed on the *collapsed and propagated* variants of the outputs

Submission	Precision	Recall	F1
Attardi run1	**81.89**	**90.45**	**85.95**
Attardi run3	81.54	90.37	85.73
Attardi run2	81.57	89.51	85.36
Mazzei run2	80.47	89.98	84.96
Lavelli run1	80.30	88.93	84.39
Mazzei run1	80.88	87.97	84.28
Lavelli run2	79.13	87.97	83.31
Grella	80.15	85.89	82.92
Lavelli run3	78.28	88.09	82.90

For the DPIE subtask, which involves an alternative, semantically–oriented evaluation, the system outputs where automatically converted to the *collapsed and propagated* notation. Table 3 reports *Precision*, *Recall* and *F1* scores achieved by each submission on the selected relations (with bolded values corresponding to the best scores).

By comparing the results in Tables 1 and 3 it can be noted that, except for the 3 top runs by Attardi et al., which remain at the top in both cases, although with a slight reordering, two runs by Mazzei (run2) and Lavelli (run1) reach higher ranks, while run3 by Lavelli drops significantly from the top ranks to the bottom rank. This sensitivity to the evaluation metrics in the ranking of system outputs was somewhat unexpected and raises the question of which evaluation metrics is more suitable to measure the performance, when aiming at Information Extraction tasks.

For the CLAP task, only one participant, Mazzei, submitted results. He focused on four languages only (Brazilian-Portuguese, French, German and Spanish) and used a single parser, the MALT parser. His strategy consisted in three steps: 1) for each language a training set was obtained by combining the Italian training set with a word–for–word translation into Italian, by using Google Translate, of the corresponding development set; 2) the best feature configuration was selected for training the parser on each language by using MaltOp-

Table 4
Results of Mazzei submissions to CLaP in terms of LAS, UAS, LA on the test sets

	LAS	UAS	LA
Brazilian-Portuguese	71.70	76.48	84.50
French	71.53	77.30	84.41
German	66.51	73.86	79.14
Spanish	72.39	77.83	83.30

timizer [8] on the combined development sets; 3) each parser was applied to a word–for–word translation into Italian of each test set. Table 4 reports, for each target language, the results of Mazzei submissions in terms of LAS, UAS and LA (Label Accuracy Score).

2.4. Discussion

For the DPIE task, the standard evaluation in terms of LAS/UAS computed on individual attachments does not seem to always correlate with the evaluation based on semantically-oriented relations, which are more relevant for Information Extraction applications, as suggested among others by [79].

A possible explanation in this case is that the technique of parser combination performed by some of the systems affected some dependency types that were relevant in extracting relations but less relevant for the attachment scores. To better understand the reported misalignment of results across the two evaluations, we performed a dependency–based analysis, focusing on single core relations and visualized in Fig. 2. Each outline corresponds to the *F1* scores associated with each relation by a participant system. It can be observed that, also in this case, there is a significant overlapping of the outlines: low scored relations are hard to predict for every participant system, although at a different extent. This suggests that either there is not enough information for dealing with semantically–oriented distinctions (as in the case of `iobj`, `npadvmod` and `tmod`), or more simply the dimension of the training corpus is not sufficient to reliably deal with them. At the same time the drop in performance for some system with respect to others is quite visible, indicating a potential for improvement.

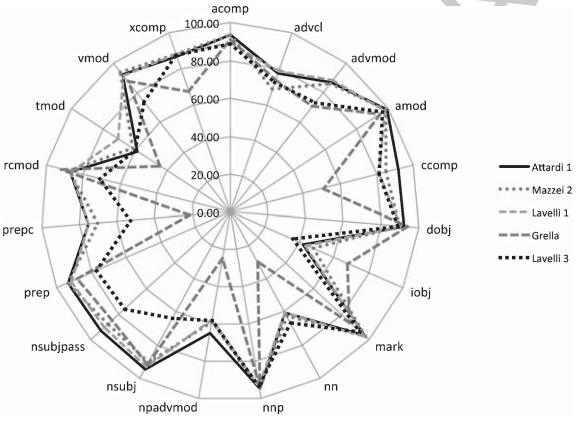

Fig. 2. Comparison on core relations.

CLAP results confirm the hypothesis behind the task, i.e. that using training data from different languages can improve accuracy of a parsing system on a given language: this can be particularly useful for improving the accuracy of parsing less–resourced languages. Given the novelty of the task, it was not possible to provide a competitive evaluation of different systems in the same task, but the results achieved are very good when compared to similar results in the literature for cross–lingual transfer parsing, where performances reported are rarely above 60% for LAS [48].

3. Evaluation of Events and Temporal Information – EVENTI

The EVENTI (*EValuation of Events aNd Temporal Information*) evaluation exercise[11] aims at promoting research in Temporal Processing in Italian[12].

Temporal Processing is an NLP task whose goal is to automatically detect and interpret events (e.g. *to walk*), temporal expressions (e.g. *May 20, 2014*) and temporal relations (e.g. in *Police arrested one suspect after a chase* the event *arrested* happened AFTER the event *chase*) within texts.

Temporal Processing is an active area of research driven by the development of annotation guidelines and international standard (i.e. TimeML [57] and ISO-TimeML [41] respectively), corpora (the English TimeBank [58] and the Spanish and Catalan TimeBanks [68] among others) and international evaluation campaigns (i.e. three editions of the TempEval task[13], the cross-document TimeLine task[14] and the QA TempEval[15], all organized in the context of SemEval workshops from 2007 up to 2015).

Despite the always increasing amount of studies and tools of Temporal Processing for various languages, previous works in Temporal Processing for Italian have been limited: a small dataset (around 30k tokens) annotated with the Italian TimeML Annotation Specifications (henceforth, It-TimeML) [24] has been released within the TempEval-2 task and only independent modules for event [26, 64] and temporal expressions [74] processing have been developed.

[11] https://sites.google.com/site/eventievalita2014/

[12] https://sites.google.com/site/ittimeml/home

[13] TempEval-1: http://www.timeml.org/tempeval/; TempEval-2 http://timeml.org/tempeval2/; TempEval-3 http://www.cs.york.ac.uk/semeval-2013/task1/

[14] http://alt.qcri.org/semeval2015/task4/

[15] http://alt.qcri.org/semeval2015/task5/

In this context, the EVENTI exercise has been designed in strict conjunction with the previous TempEval campaigns and standard annotation initiatives with the purpose of advancing the state of the art of Temporal Processing for Italian by proposing, for the first time, a Main task on contemporary news and a Pilot task on historical texts.

3.1. Data: annotation and preparation

The EVENTI exercise is based on the EVENTI Annotation Guidelines[16], a simplified version of the It-TimeML guidelines that are compliant with the TimeML [57] and ISO-TimeML [41] Annotation Guidelines.

The EVENTI Annotation Guidelines identify four types of temporal information to be annotated: temporal expressions (timexes), events, signals and temporal relations (TLINKs).

A timex is a single token or a sequence of tokens with a temporal meaning. A timex may be realized by adverbs (e.g. *ieri* [yesterday]), nouns (e.g. *alba* [dawn]), numerical expressions (e.g. *1980*), adjectives (e.g. *annuale* [yearly]) and calendar expressions (e.g. *24-07-2014*). All timexes are marked with the <TIMEX3> tag, assigned with a type (i.e. *DATE*, *TIME*, *DURATION*, *SET*) and normalized giving them a value (e.g. *eleven in the morning* has a value of 'T11:00'). When a timex is referenced indirectly in a text, an empty, non-text consuming <TIMEX3> tag, i.e. a tag with no textual extent, is created. For instance, in *dalle due alle cinque del pomeriggio* [from two to five in the afternoon] the implicit duration between *due* [two] and *cinque* [five] must be represented with a non-consuming tag expressing a value of three hours (i.e. <TIMEX3 value ="P3TH"/>).

An event is anything that happens, occurs or a circumstance in which something holds true. Such broad notion of event corresponds to what in the literature is called *eventuality* [7]. A wide range of linguistic expressions may realize events, such as tensed and untensed verbs, nouns and nominalizations (e.g. *(ha) comprato* [(he has) bought], *correre* [to run], *distruzione* [destruction], *pace* [peace]). All events are annotated by marking their extent with the <EVENT> tag and by assigning values to 12 different attributes which contribute to make explicit both morphosyntactic and temporal information of each eventuality. In particular, the CLASS attribute allows to classify eventualities following semantic and syntactic criteria so to facilitate temporal and factual reasoning [66].

A signal is a linguistic element, such as a preposition or a conjunction, which either directly or indirectly suggests the presence of a temporal relation: e.g. *durante* [during], *dopo* [after]. They are marked up using the <SIGNAL> tag.

A temporal relation is a link between two annotated elements (events or timexes) that are temporally related to each other. As the EVENTI exercise is the first complete evaluation exercise for temporal processing in Italian, the annotation of temporal relations was limited to a set of 3 cases so as not to over-complicate an already complex task: (i) relations between an eventuality and a timex in the same sentence (e.g. between the eventuality *venduto* [sold] and the timex *1983* in the sentence *ha venduto 3 milioni di lavatrici nel 1983* [it sold 3 millions washing machines in 1983]); (ii) relations between a main event (i.e. the syntactic root of the sentence) and a subordinated event in the same sentence (e.g. between *dichiarato* and *rilasciati* in *ha dichiarato che gli ostaggi sono stati rilasciati* [he declared that the hostages have been released]); (iii) relations between main events in coordinated clauses (e.g. between *ascoltato* [listened] and *risposto* [answered] in *aveva ascoltato con attenzione e risposto ad ogni domanda* [he had carefully listened and answered every question]). All temporal relations are marked with the <TLINK> tag and classified with respect to a set of 13 values based on Allen's interval relations [1] (e.g. AFTER and its inverse relation BEFORE).

Following the EVENTI Annotation Specifications, 3 datasets in stand-off XML have been annotated: the Main task training data, the Main task test data and the Pilot task test data. The news stories for the Main task are taken from the Ita-TimeBank [25]. Two expert annotators conducted a revision of the annotations for the Main task to solve inconsistencies[17] using CAT[18] [9]. The final size of the EVENTI corpus for the Main task is of 130,279 tokens, divided in 103,593 tokens for training and 26,686 for test. The Main task training data have been released to participants in two separate batches[19] and have a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 license. The Pilot test data consist of about 5,000 tokens of historical texts, i.e.

[16] http://goo.gl/IgO6A1

[17] Readers are referred to the paper describing the Ita-TimeBank for details on the inter-annotator agreement.

[18] http://dh.fbk.eu/resources/ cat-content-annotation-tool

[19] ILC Training Set: http://goo.gl/3kPJkM; FBK Training Set: http://goo.gl/YnQWml

Table 5
Annotated events, temporal expressions, signals and temporal relations in the EVENTI corpus

Tag	Main training	Main test	Pilot test
EVENTs	17,835	3,798	1,195
TIMEX3s	2,735	624	97
SIGNALs	932	231	62
TLINKs	3,500	1,061	382

articles dating back to 1914 and related to the outbreak of World War I. They have been manually annotated in CAT by an expert annotator. For the Pilot task, no training data have been provided and participants were asked to re-run the systems developed for the Main task. Table 5 reports the total number of annotated elements and links in the 3 datasets composing the EVENTI corpus.

3.2. Subtask description

Main and Pilot tasks are composed by the following subtasks.

Subtask A: To determine the extent, type and value of timexes according to the TIMEX3 tag specifications. For the first time, empty TIMEX3 tags are taken into account in the evaluation.

Subtask B: To determine extent and class of eventualities in a text according to the EVENT tag specifications.

Subtask C: To identify temporal relations in raw texts. This subtask implies performing subtasks A and B and subsequently identifying the pairs of elements (event - event and event - timex pairs) which stand in a temporal relation (TLINK) and classifying the temporal relation itself. The set of candidate pairs has been limited to the cases reported in Section 3.1. All temporal relation values in It-TimeML are used.

Subtask D: To determine the value of the temporal relation given two gold temporal elements (i.e. the source and the target of the relation) as defined in Task C.

The evaluation measures of the EVENTI exercise are a re-implementation of the measures used in the TempEval-3 task [77]. In particular, the scorer was adapted to the CAT XML format and the evaluation of temporal expressions was extended to include empty TIMEX3 tags. For subtask A and B, both the correctness of tags' extent and that of attributes' values were evaluated. More specifically, for subtask A the correctness of the normalized value and for subtask B the correctness of the class value were evaluated. For recognition, Precision, Recall and F1-score were used. As for attribute evaluation, F1-score to measure how well a system identifies an element and its attribute values was used. As for subtask C, three aspects were evaluated: i.) the number and the extent of the temporal elements identified in a raw text; ii.) the identification of the correct sources and targets; and, iii.) the identification of the correct temporal value. In subtask D, only the identification of the correct temporal value was evaluated. Similarly to subtasks A and B, Precision, Recall and F1-score were also computed for subtasks C and D. Rankings for all subtask are based on F1-scores. In particular, the final ranking is based on the value attribute F1 score for subtask A (temporal expression recognition and normalization), on the class attribute F1 score for subtask B (event detection and normalisation), on the "overall temporal awareness" of the system for subtask C (temporal relation identification and classification from raw text) and, finally, on the temporal value attribute F1 for subatsk D (temporal relation classification given gold entities)

3.3. Participant systems and results

Three teams submitted system results for a total of 17 unique runs: FBK (Fondazione Bruno Kessler), HT (University of Heidelberg), and UNIPI (Università di Pisa).

FBK is an end-to-end system based on a machine learning approach. It combines and adapts to Italian three subsystems on finally developed for English: one for timex recognition and normalization, one for event extraction, and one for temporal relation identification and classification. Timex normalisation is conducted by means of an adaptation to Italian of TimeNorm [12], a rule-based system based on synchronous context free grammars. The other subsystems are based on Support Vector Machine approaches. HeidelTime (HT) is a rule-based, multilingual and cross-domain temporal tagger [73]. UNIPI used the publicly available version of HeidelTime and adapted it by integrating into the pipeline the Tanl tools [5].

FBK took part to all subtasks, while HT and UNIPI only to Subtask A. Table 6 reports the best scores for the EVENTI subtask. Next to each result, the acronym of the best system is reported. In bold the winning system. Readers are referred to the proceedings of EVALITA 2014 for detailed descriptions of the systems [21].

Table 6
Comparison of the best results obtained in the Pilot and in the Main tasks of the EVENTI evaluation exercise, including the acronyms of the best systems

	F-Measure Best System		
	Sub task	Main task	Pilot task
A	TIMEX recognition	0.886 (FBK)	0.870 (FBK)
	TIMEX normalization	**0.709 (HT)**	**0.475 (FBK)**
B	EVENT detection	0.884 (FBK)	0.843 (FBK)
	EVENT classification	**0.671 (FBK)**	**0.604 (FBK)**
C	TLINKs raw text	**0.264 (FBK)**	**0.185 (FBK)**
D	TLINKs gold items	**0.736 (FBK)**	**0.419 (FBK)**

3.4. Discussion

The results suggest observations on strong and weak aspects of temporal processing.

Subtask A has shown how for specific tasks rule-based approaches are competitive with statistical ones (e.g. timex recognition), if not the best solution (e.g. timex normalization). In subtask B, event detection has reached good results. In this case, provided the complex interplay of various linguistic elements in the identification of an event, the statistical approach seems to work best. The lower performances in event classification can be hardly explained, as FBK is the only participant system and the possible sources of errors ranges from errors in annotation, lack of data in the training data and issues in the selected approach. On event classification, the winning strategy adopted by FBK was to approach the task in two steps: first, event detection, and then, use the results of this subtask as features to model the classification approach. An interesting aspect is the use of lexical resources, such as MultiWordNet [55], to include lexical semantics information. Although no direct comparison could be conducted with previous systems, due to the fact that the original TempEval-2 Italian test set has been included in the EVENTI training, the figures reported in [26] suggest that finer grained lexical knowledge for event classification may improve the results.

The most interesting results are those concerning subtask C. Subtask C aims at evaluating the temporal awareness of a system. The results obtained are not satisfactory, although the subset of temporal relations was carefully selected and formalized. Identifying the relevant elements which enter in a temporal relation is not a trivial task.

A downgrading of the results in the Pilot dataset can be observed for all subtasks. This is due to the different genre of the texts composing the dataset. Historical texts vary on different levels with respect to contemporary newspaper articles. For instance, lots of fuzzy timexes which require special rules and are not present in the training data, such as *il giorno (della vittoria)* [the day (of the victory)], were identified. Similar observations hold also for events. The lack or sparseness of positive examples, namely for complex events such as those realized by nouns, is the major reason for the lower performances in subtask B.

One of the major achievement of the EVENTI campaign is the development of end-to-end systems and the assessment of state-of-the-art results for temporal processing in Italian. As already stated, comparison with previously developed modules cannot be conducted due to differences in the dataset and evaluation measures adopted. A tentative comparison with the results obtained in other languages, such as English and Spanish, can be carefully conducted. Concerning subtask A (temporal expressions recognition and normalisation) and subtask B (event detection and classification) the results in Italian are satisfying and in line with those reported for English (subtask A F1 0.776 and subtask B F1 0.718) and Spanish (subtask A F1 0.875 and subtask B 0.576) at TempEval-3 [77]. The lower performance of Italian on subtask A can be due to the inclusion of inferred temporal expressions in the evaluation, excluded from TempEval-3. More interesting results emerge for subtasks C and D. Although full temporal processing has been simplified in Italian, by limiting the set of possible entities and temporal relations, the low performance on Italian data (0.264) is pretty similar to the best scores for English (F1 0.309) and Spanish (F1 0.416), suggesting that the task is really hard and different solutions and approaches should be engaged. Finally, in subtask D the Italian system obtains really good results with respect to English (F1 0.564)[20].

4. Sentiment Polarity Classification – SENTIPOLC

The main goal of the SENTIPOLC shared task (SENTIment POLarity Classification) is sentiment analysis at the message level on Italian tweets.

Sentiment analysis (SA), which has been defined as "the computational study of opinions, sentiments and emotions expressed in text" [14], has become a relevant topic of research in NLP, especially with respect to the study of new forms of digital and social communication. The huge amount of information streaming from online social networking and micro-blogging platforms

[20] No system was submitted for Spanish.

such as Twitter is increasingly attracting the attention of many kinds of researchers and practitioners. The fact that, as in 2013 [50], also in 2014 *Sentiment Analysis in English tweets* was the most popular Semeval shared task, with over 40 participating teams for the subtask B (message-level sentiment analysis) is indicative in itself [65]. SENTIPOLC was the most participated task of EVALITA 2014 with a total of 35 submitted runs from 11 different teams, reflecting the great interest of the NLP community in sentiment analysis in social media, also in Italy.

As highlighted by Bo Pang and Lillian Lee in their seminal paper, the terms *opinion mining* and *sentiment analysis* are often used interchangeably to denote the same field of study: "when broad interpretations are applied, 'sentiment analysis' and 'opinion mining' denote the same field of study (which itself can be considered a sub-area of subjectivity analysis)[21]" [53]. In [23] Cambria and colleagues point out how "opinion mining and sentiment analysis actually focus on polarity detection and emotion recognition, respectively", but since the two aspects are usually interrelated they are used as synonyms.

Sentiment analysis can be articulated in related tasks, with slightly different focuses. The basic task is *polarity classification*, which occurs when a piece of text stating an opinion on a given target is classified as expressing one of two opposing sentiments: systems need to understand the positive or negative sentiments in the text, when present. Various issues should be taken into account. First, texts can contain parts expressing mixed sentiment (perhaps because they comment on more than one topic), a feature that should be tackled. Second, some texts don't contain strong opinions, but have a clearly subjective flavor. In this respect, it is interesting to consider the related *subjectivity detection* task, which aims to distinguish between subjective and objective texts, and can possibly help in classifying the sentiment. Furthermore, the fact that social communications include a high percentage of ironic messages cannot be neglected [33, 47, 63]. In order to investigate this issue, a pilot subtask on *irony detection* has been included in the battery of SENTIPOLC subtasks.

Sentiment analysis naturally relies on resources such as sentiment annotated datasets, sentiment lexica (i.e. resources such as dictionaries or lists of words labeled with sentiment polarity), and the like. However, the availability of resources for languages other than English is usually rather scarce, and this holds for Italian as well [11, 19]. The organisation of the SENTIPOLC shared task was thus aimed at providing reliably annotated data as well as promoting the development of systems towards a better understanding and processing of how sentiment is conveyed in tweets.

4.1. Task description

The SENTIPOLC shared task comprises three subtasks of increasing complexity. Participants could choose to take part in one or more subtasks.

Task 1: Subjectivity Classification: *decide whether a given message is subjective or objective.* This is a standard task on recognising whether a message is subjective or objective.

Task 2: Polarity Classification: *decide whether a given message is of positive, negative, neutral or mixed sentiment.* Differently from most SA tasks, chiefly the Semeval 2013 task, in our data positive and negative polarities are *not* mutually exclusive. This means that a tweet can be at the same time positive *and* negative, yielding a *mixed* polarity, or also neither positive nor negative, meaning it is a subjective statement with *neutral* polarity[22].

Task 3 (Pilot): Irony Detection: *decide whether a given message is ironic or not.* Twitter communications include a high percentage of ironic messages [33, 47, 63], and platforms monitoring the sentiment in Twitter messages experienced the phenomenon of wrong polarity classification in ironic messages [19]. Indeed, the presence of ironic devices in a text can work as an unexpected "polarity reverser" (one says something "good" to mean something "bad"), thus undermining systems' accuracy.

The three tasks are meant to be completely independent. For example, a team could take part in the polarity classification task, which only applies to subjective tweets, without tackling Task 1. For each task, each team had to submit a *constrained* run, using the provided training data only (and resources such as lexicons, but not additional training data) and optionally an *unconstrained* run where using additional data for training, i.e. more sentiment annotated tweets, is allowed.

[21] Subjectivity analysis is here meant as dealing with the recognition of opinion-oriented language in order to distinguish it from objective language.

[22] In accordance with [80].

4.2. Data

The data exploited for this shared task is a collection of tweets derived from two existing corpora, namely SENTI-TUT [19] and TWITA [11]. Both corpora have been revised according to the new annotation guidelines specifically devised for this task.

There are two main components of the dataset: a *generic* and a *political* collection. The latter has been extracted exploiting specific keywords and hashtags marking political topics, while the former is composed of random tweets on any topic. Each tweet is thus also marked with a "topic" tag.

A tweet is represented as a record containing the Twitter id, the fields representing subjectivity, positive polarity, negative polarity and irony, and the topic field. Apart from the id, the fields are binary values, encoding the absence/presence of the feature. For the topic field, 0 means "generic" and 1 means "political". The fields with manually annotated values are: `subj`, `pos`, `neg`, `iro`. While these classes could be in principle independent of each other, the following constraints hold in the annotation scheme:

- An objective tweet will not have any polarity nor irony, thus if `subj` $= 0$, then `pos` $= 0$, `neg` $= 0$, and `iro` $= 0$.
- A subjective tweet can exhibit at the same time positive *and* negative polarity (mixed), thus `pos` $=$ 1 and `neg` $= 1$ can co-exist.
- A subjective tweet can exhibit no specific polarity and be just neutral but with a clear subjective flavour, thus `subj` $= 1$ and `pos` $= 0$ and `neg` $=$ 0 is a possible combination.
- An ironic tweet is always subjective and it must have a defined polarity, i.e. `iro` $= 1$ cannot be combined with `pos = neg`.

According to this schema, seven possible combinations of labels are allowed. To give a couple of examples, this is an objective tweet (1000) from the SENTIPOLC dataset: *l'articolo di Roberto Ciccarelli dal Manifesto di oggi* http:// fb.me/1BQVy5WAk [The article by Roberto Ceccarelli from today's Manifesto [URL]]; this is instead a subjective tweet with negative polarity, and an ironic twist (1011): *Botta di ottimismo a #lInfedele: Governo Monti, o la va o la spacca* [Optimism boost at #lInfedele: Monti administration, make it or break it].

The SENTI-TUT section of the dataset was previously annotated for polarity and irony[23]. The tags POS, NEG, MIXED and NONE[24] in Senti–TUT were automatically mapped in the following values for the SENTIPOLC's `subj`, `pos`, `neg`, and `iro` annotation fields: POS $\Rightarrow$ 1100; NEG $\Rightarrow$ 1010; MIXED $\Rightarrow$ 1110; NONE $\Rightarrow$ 0000. However, the original Senti–TUT annotation scheme did only partially match the one proposed for this task, in particular regarding the ironic tweets, which were annotated just as HUM in SENTI–TUT, without polarity. Thus, for each tweet tagged as HUM ($\sim$800 tweets), two annotators independently added the polarity dimension. The inter-annotator agreement at this stage was low: $\kappa = 0.259$; as expected, the presence of irony affects the perceived sentiment of a tweet, by introducing a further element of subjectivity making it more difficult to find an agreement among human annotators. In a second round, a third annotator attempted to solve the disagreements ($\sim$33%). Tweets where all three annotators had a different opinion ($\sim$10%) were discussed jointly for the final label assignment.

The TWITA section of the dataset had to be completely re-annotated, as irony annotation was missing, and the three labels adopted in the original data were not directly transferrable to the new scheme see [11]. The annotation was performed by four experts in three rounds. Round one saw two annotators independently mark each tweet. Inter-annotator agreement was measured at $\kappa = 0.482$ for Task 1, $\kappa = 0.678$ for positive labels and $\kappa = 0.638$ for negative labels in Task 2, and at $\kappa = 0.353$ for Task 3. In round two, a third annotator made a decision on the disagreements from round one, and in round three a fourth annotator had to decide on those cases where disagreements were left by the previous two rounds. Tweets where all four annotators had a different opinion amounted to just nine cases, and were discussed jointly for the final label assignment.

Participants were provided with a development set (SentiDevSet henceforth), consisting of 4,513 tweets. The dataset is the same for all three subtasks.

Due to Twitter's privacy policy, tweets cannot be distributed directly, so participants were also provided with a web interface based on the use of RESTful Web API technology, through which they could download the tweet's text on the fly for all the ids provided[25].

[23] For the annotation process and inter-annotator agreement for SENTI–TUT see [19, 20].

[24] Four annotators collectively reconsidered the set of tweets tagged NONE in order to distinguish the few cases of subjective, neutral, not-ironic tweets (1000).

[25] http://www.di.unito.it/~tutreeb/sentipolc-evalita14/tweet.html.

However, some tweets for which ids were distributed, might be not available anymore at download time for various reasons: Twitter users can delete their own posts anytime; their accounts can be temporarily suspended or deactivated. As a consequence, it is possible that the number of the available messages in the development dataset will vary over time. In order to deal with this issue, at submission time participants were asked to equip their runs with the information about the number of tweets actually retrieved from SentiDevSet.

4.3. Evaluation

Task1: subjectivity classification Systems have been evaluated on the assignment of a 0 or 1 value to the subjectivity field. A response is considered plainly correct or wrong when compared to the gold standard annotation. Precision, Recall and F1-score were computed for each class (`subj`, `obj`). The systems are then ranked based on the average of the F1-scores for subjective and objective classes.

Task2: polarity classification The SENTIPOLC coding system allows for four combinations of *positive* and *negative* values: 10 (positive polarity), 01 (negative polarity), 11 (mixed polarity), 00 (no polarity). Accordingly, positive polarity and negative polarity have been evaluated independently by computing precision, recall and F1-score for both classes (0 and 1).

The F1-score for each of the two polarity classes was computed as the average of the F1-scores of their respective pairs of classes, and finally the systems were ranked according to the average of the F1-scores of the two polarities classes.

Task3: irony detection Systems are evaluated on their assignment of a 0 or 1 value to the irony field. A response is considered fully correct or wrong when compared to the gold standard annotation. We measure Precision, Recall and F1-score for each class (ironic, non-ironic). The systems are ranked based on the average of the F1-scores for ironic and non-ironic classes.

4.4. Participants and results

A total of 11 teams from four different countries participated in at least one of the three tasks of SENTIPOLC. Almost all teams participated to both subjectivity and polarity classification subtasks. Most of the submissions were constrained: 9 out of 12 for subjectivity classification; 11 out of 14 for polarity classification; 7 out of 9 for irony detection. In particular, three teams participated with both types of runs. Unconstrained systems did not show to improve performance, but actually decreased it, with one exception (UNITOR's systems).

Table 7
Task 1, subjectivity detection: F1-scores for constrained (C) and unconstrained runs (U)

rank	team	F(C)	F(U)
1	uniba2930	0.7140	0.6892
2	UNITOR	0.6871	0.6897
3	IRADABE	0.6706	0.6464
4	UPFtaln	0.6497	–
5	ficlit+cs@unibo	0.5972	–
6	mind	0.5901	–
7	SVMSLU	0.5825	–
8	fbkshelldkm	0.5593	–
9	itagetaruns	0.5224	–
10	baseline	0.4005	–

Because of the downloading procedure implemented to comply to Twitter's policies, not all teams necessarily tested their systems on the same set of tweets. Differences turned out to be minimal, but to ensure fairness evaluation was performed over an identical dataset for all. All participant systems were evaluated on the union of their classified tweets, which amounted to 1,734 tweets (1,930-196)[26].

A single-ranking table was produced for each subtask, where unconstrained runs are properly marked. Notice that only the average F1-score was used for global scoring and ranking. However, systems that are ranked midway might have excelled in precision for a given class or scored very bad in recall for another. Detailed scores for all classes and all tasks are available in the task report [10]. For each task, we ran a majority class baseline to set a lower-bound for performance. In the tables it is always reported as *baseline*.

Table 7 shows results for the subjectivity classification task. All participant systems show an improvement over the baseline.

Table 8 shows results for the polarity classification task, which was the most popular subtask. Also in this case, all participant systems show an improvement over the baseline[27].

Table 9 shows results for the irony detection task. While all participant systems show an improvement

[26] It turned out that five of the 1,935 tweets in SentiTestSet were duplicates.

[27] Itanlp-wafi submitted a new run after the deadline to correct a format error. Official ranking was not revised, but the evaluation of the correct run is shown in the table.

Table 8
Task 2, polarity detection: F1-scores for constrained (C) and unconstrained runs (U)

rank	team	F1(C)	F1(U)
1	uniba2930	0.6771	0.6638
2	IRADABE	0.6347	0.6108
3	CoLingLab	0.6312	–
4	UNITOR	0.6299	0.6546
5	UPFtaln	0.6049	–
6	SVMSLU	0.6026	–
7	ficlit+cs@unibo	0.5980	–
8	fbkshelldkm	0.5626	–
9	mind	0.5342	–
10	itagetaruns	0.5181	–
11	Itanlp-wafi*	0.5086	–
12	baseline	0.3718	–
	*amended run	0.6637	–

Table 9
Task 3, irony detection: F1-scores for constrained (C) and unconstrained runs (U)

rank	team	F1(C)	F1(U)
1	UNITOR	0.5759	0.5959
2	IRADABE	0.5415	0.5513
3	SVMSLU	0.5394	–
4	itagetaruns	0.4929	–
5	mind	0.4771	–
6	fbkshelldkm	0.4707	–
7	UPFtaln	0.4687	–
8	baseline	0.4441	

over the baseline, this time some systems score very close to it, highlighting the complexity of the task.

4.5. Discussion and conclusions

Participant systems were compared according to the following main dimensions: exploitation of further annotated data for training, classification framework (approaches, algorithms, features), exploitation of available resources (e.g. sentiment lexicons, etc.), issues about the interdependency of tasks.

Most participants submitted constrained systems only. Three teams submitted unconstrained runs, and apart from UNITOR, results are worse than those obtained by the constrained runs. Likely this situation is triggered by the current lack of sentiment-annotated, available large datasets for Italian. Additionally, what might be available is not necessarily annotated according to the same principles adopted in SENTIPOLC. Interestingly, uniba2930 attempted acquiring more training data via co-training. They trained two SVM models on SentiDevSet, each with a separate feature set, and then used them to label a large amount of acquired unlabelled data progressively adding training instances to one another's training set, and re-training. No significant improvement was observed, due to the noise introduced by the automatically labelled training instances.

As noticed also in the context of similar evaluation campaigns for English [50, 65], most systems used supervised learning, the exceptions being itagetaruns and ficlit+cs@unibo. The most popular algorithm was SVM, but also Decision Trees, Naive Bayes, K-Nearest Neighbors were used. As mentioned, one team experimented with a co-training approach, too.

A variety of features were used, including word-based, syntactic and semantic (mostly lexicon-based) features. The best team in Task1 and Task2, uniba2930, specifically mentions that in leave-one-out experiments, (distributional) semantic features appear to contribute the most. uniba2930 is also the only team that explicitly reports using the topic information as a feature, for their constrained runs. The best team in Task3, UNITOR, employs two sets of features explicitly tailored for the detection of irony, based on emoticons/punctuation and a word space model to identify words that are out of context. Typical Twitter features were also generally used, such as emoticons, links, usernames, hashtags.

Two participants did not adopt a learning approach. ficlit+cs@unibo developed a system based on a sentiment lexicon that uses the polarity of each word in the tweet and the idea of "polarity intensifiers". A syntactic parser was also used to account for polarity inversion cases such as negations. itgetaruns was the only system solely based on deep linguistic analysis exploiting rhetorical relations and pragmatic insights.

Almost all participants relied on various sentiment lexicons. At least six teams used information from SentiWordNet [37], either using the already existing Sentix [11] or otherwise. Several other lexica and dictionaries were used, either natively in Italian or translated from English (e.g. AFINN, Hu-Liu lexicon, Whissel's Dictionary). Native tools for Italian were used for pre-processing, such as tokenisers, POS-taggers, and parsers.

The majority of systems participating in more than one subtask adopted classification strategies including some form of interdependency among the tasks, with different directions of dependency.

From a first comparative analysis of the systems' behaviour some observations can be done, related to aspects specific to the SENTIPOLC tasks. First, ironic expressions do appear to play the role of polarity

reversers, undermining the accuracy of sentiment classifiers. Second, recognising mixed sentiment (tweets tagged as 1110) was hard for our participants, even harder than recognising neutral subjectivity (tweets tagged as 1000). Finally, recognizing positive sentiment was hard for our participants while systems were better in recognizing the presence of negative sentiment (and the absence of positive sentiment). This can be due to several reasons: the SENTIPOLC corpus (maybe also because of the presence of tweets on politics) has a natural bias towards negative sentiment, and also sentiment lexicons available for this task seem to have a bias towards negative words.

5. Applicative impact of shared tasks results

In this section we will report some conclusions by highlighting the potential impact of shared tasks results on real world applications and further directions of research which emerged from their analysis.

Thanks to its simple and direct encoding of predicate–argument structures, dependency parsing is an attractive technique for use in applications such as Information Extraction, Question Answering, Machine Translation, Language Modelling, Semantic Role Labelling, and Textual Entailment. Practical uses of text analysis based on dependency structures are reported in many applications and domains, including medical, financial or intelligence. Google, for example, applies dependency parsing to most texts it processes [38]: parse trees are used in extracting relations to build the Knowledge Vault [35] and to guide translation [42]. Its central role in a wide range of applications makes dependency parsing a still interesting task to be explored in evaluation campaigns.

In this EVALITA edition, most participant systems resorted to the combination of two or more parsers rather than to a single system. Such an approach exploits the fact that the parsers differ in their strengths and error types and for this reason systems relying on combination approaches yield higher results. However, if parser output combination is understandable when aiming to achieve the best possible accuracy, it is less desirable when one expects to use the parser in a production setting, i.e. where it must be embedded in a larger pipeline and considerations of speed and memory usage become relevant. In spite of their high accuracy, combination approaches are not expected to be suitable for large scale applications due to their computational cost in terms of memory and speed. From this it follows that the focus should rather be on on improving single parsers technology, by working in several directions. For instance, by incorporating semantic knowledge, e.g. distributed semantic representations, in order to improve the accuracy on semantic attachments [28]; or by exploiting alternative sets of parsing rules [38, 67]; or by looking at dependency parsing as part of a wider task also including morphological analysis, Part–of–Speech tagging and Semantic Role Labeling. In spite of focusing on the accuracy of the participant systems only, future dependency parsing shared tasks should try to balance the different and contrasting requirements of accuracy and performance.

Performance and accuracy of parsing systems are not the only concerns which need to be tackled from this perspective. As pointed out in Section 2, the standard evaluation of dependency parsing in terms of LAS/UAS computed against individual attachments does not seem to always correlate with the evaluation based on semantically–loaded relations, which play a central role in Information Extraction applications. This is an open issue which applies to the evaluation of individual systems and which needs to be addressed in future evaluation campaigns. A non marginal aspect related to this issue concerns the evaluation measures which should be devised to reflect more closely the impact of the achieved results on more complex processes such as Information Extraction.

Another aspect worth mentioning as a positive outcome of the EVALITA 2014 shared task on dependency parsing is the resource ISDT itself, which is the largest Italian dependency treebank compliant with international standards. Thanks to the work done in the Stanford Dependencies framework to build ISDT, it was possible to join, since the beginning, the Universal Dependencies project [32], a large international standardization endeavour involving already 20 different languages, which has a great potential for fostering the development of cross–language analysis tools.

The EVENTI (*EValuation of Events aNd Temporal Information*) evaluation exercise evolves around temporal processing, which plays a central role in a wide number of NLP applications. For example, Question Answering needs to resolve temporal expressions and automatically identify events in order to answer questions about when a particular event occurs or which events occur in a temporal relation to another given event [59]. Temporal awareness is important also for Multi-Document Summarization in order to avoid inappropriate merging of distinct events [29]. In such a cross-document and cross-temporal perspective,

temporal information is the basic building block for the development of more complex systems which aims at reconstructing storylines, i.e. the plot of the "story" of a target entity. Storyline extraction can have a relevant impact also in other domains such as policy making or industries. Being able to collect relevant events for specific entities can allow to monitor past activities, opinions, promises, contradicting information, and, possibly, predict future trends[28]. The impact of temporal information also affects Information Extraction systems. Assertions about entities and relations are incomplete or incorrect if it is impossible to capture how their properties are temporally updated [3].

So far, only the news domain and, more recently, the clinical domain [69] have been extensively investigated. Nevertheless, the recognition and elaboration of temporal information is a crucial step when dealing with history-related matters. Beyond the need to support historical investigation with computational approaches for the specific purposes of historians, there is a more general requirement, i.e. to enhance historical research to improve human understanding of the past. History is life's teacher (Cicero in *Pro Publio Sestio* oration, "History is the witness that testifies to the passing of time; it illumines reality, vitalizes memory, provides guidance in daily life and brings us tidings of antiquity") thus digital tools that enhance historical research could improve human knowledge of the past supplying useful information to politicians and citizens to understand both the present and the future of our society[29].

For what concerns SENTIPOLC, analyzing sentiment and opinions in social media is a theme of great interest for industry, in several respects. Many companies are interested in using results of opinion mining and sentiment analysis in order to develop marketing strategies. In fact, user-generated contents, such as tweets, are a precious mine for grasping opinions of people about a specific topic or product, thus, they can constitute a valuable asset for firms to directly tap into the customer's preferences. But the leveraging of social media for the purpose of tracking product image requires sentiment-related technologies, and in response to this needs NLP related companies that offer monitoring and analysis of social media to learn more about consumer behavior towards brands, products and services are getting more and more popular. On this line, forums such as the American *Sentiment Analysis Symposium*[30] are annually organized with the explicit aim to bridge business and sentiment-related technology for mining and exploiting opinions, emotions, and intents in online, social, and enterprise contents.

However, prospective applications of sentiment analysis are not limited to business services for companies, but can be envisioned across different domains ranging from politics to sociology. Opinion mining of data extracted from social media can be used in the public sector, in order to introduce an e-participation perspective into the policy making life cycle. Policymakers indeed need advanced e-participation tools for being supported in their work, both at the decision-making stage, and in the ex-post evaluation of the impact of their policies (see e.g. the e-Policy EU Project[31]). Sentiment analysis in Twitter has been also used to monitor political sentiment [76] and to extract critical information during times of mass emergency [78]. New areas of research have also arisen in the social sciences field, such as the Subjective Well-Being in Psychology [34] and the Happiness economics in Economy, within the debate on alternative measure to Gross Domestic Product. Here, sentiment analysis could contribute to interpret the degree of well-being of a country. In some pioneering works in this direction, extracting sentiments from Twitter data has been used to detect moods and happiness in a given geographical area [2, 49], to look for correlation between mood and traditional economic indicators [14], and to measure the well-being of a population [62].

The SENTIPOLC shared task was open to everyone from industry and academia, and the organizing team included among its members a representative of CELI, one of the Italian companies interested in the SENTIPOLC topic, as providing services of monitoring and analysis of social media based on the exploitation of NLP technologies. However, only few non-academic organizations declared interest in the task, asking for test data, and in the end only academic teams, from different countries, submitted their run for the evaluation. The discussion at the workshop with the participant teams was focussed on lessons learned and feedback for a future edition of the task. Interesting directions of research include: aspect–based sentiment analysis,

[28] EU NewsReader project (FP7-ICT-2011-8 grant agreement no.: 316404) and NWO Spinoza Prize project Understanding Language by Machines (subtrack 3).

[29] "Tools for synthesising information about change over time are of increasing importance in an era marked by a crisis about the future, when most institutions do their planning on cycles of less than five years"[39].

[30] http://sentimentsymposium.com/

[31] e-Policy - Engineering the POlicy-making LIfe CYcle: http://www.epolicy-project.eu/

with a specific focus on the *target* of the sentiment expressed in the tweets, as already done in the framework of Semeval 2014 [56]; fine-grained categorization with respect to the type of affective state to detect, by moving towards a task of classification of tweets in different emotion categories (e.g. Ekman's emotions like joy, fear, sadness, anger, disgust, or surprise), which represents a challenging but not trivial task to organize, especially because of the difficulties related to the development of a reliable gold standard [72].

The three shared tasks selected for the 2014 edition of EVALITA are innovative *per se* and individually show an enormous potential for real world applications. However, when looked at from a wider perspective they present valuable and bidirectional interconnections which are worth being emphasized here. On the one hand, dependency parsing represents a key step of analysis from which other tasks, such as temporal processing and sentiment analysis, can definitely benefit. From the opposite perspective, the SENTIPOLC and the EVENTI tasks represent additional challenges for dependency parsing algorithms, which need to be adapted to reliably deal with non–canonical varieties of language, going from the language of social media and computer–mediated communication in general to historical data.

The interconnections among tasks, however, are not restricted to whether and to what extent an individual task can contribute to the improvement of another. On the applicative front, more challenging and complex tasks can be envisaged on the basis of current datasets from this and previous EVALITA editions: this is the case, for example, of a composite and articulated shared task aimed at tracking sentiment across time with respect to a given target (be it an entity or an event) where dependency parsing represents a basic and unavoidable pre–processing step. This cross–fertilization among different tasks can lead to new, more challenging and articulated shared tasks which could be considered for the organization of future EVALITA editions. An Information Extraction process can be very complex and for this reason it is commonly decomposed into distinct basic tasks. Future EVALITA shared tasks could be framed in this wider perspective: i.e. they could be seen as individual components of a wider Information Extraction architecture where evaluation of participant systems could be carried out at the level of both the individual tasks and the wider Information Extraction process.

The overall experience of the 2014 edition of EVALITA showed that language technology research on Italian is vibrant in several directions. Together with the consolidation of the research community focused on Italian, the results gathered from the organization of the evaluation campaign help to bring new perspectives towards old and new challenges in the field.

References

[1] J.F. Allen, Towards a general theory of action and time, *Artificial Intelligence* **23**(2) (1984), 123–154.

[2] L. Allisio, V. Mussa, C. Bosco, V. Patti and G. Ruffo, Felicittá: Visualizing and estimating happiness in Italian cities from geotagged Tweets, *in Proceedings of the 1st International Work-shop on Emotion and Sentiment in Social and Expressive Media (ESSEM@AI*IA), CEUR-WS.org*, volume 1096, 2013, pp. 95–106.

[3] O. Alonso, K. Berberich, S.J. Bedathur and G. Weikum, NEAT: News Exploration Along Time, *in Proceedings of the 32nd European Conference on Information Retrieval (ECIR 2010), Lecture Notes in Computer Science 5993*, Springer, Berlin/Heidelberg, 2010, p. 667.

[4] G. Attardi, Experiments with a Multilanguage Non-Projective Dependency Parser, *in Proceedings of the 10th Conference on Natural Language Learning, ACL*, New York, NY, 2006, pp. 166–170.

[5] G. Attardi, S. Dei Rossi and M. Simi, The Tanl Pipeline, *in Proceedings of the LREC Workshop on Web Services and Processing Pipelines in HLT (WSPP), ELRA*, 2010, pp. 15–21.

[6] G. Attardi and M. Simi, Dependency Parsing Techniques for Information Extraction, *in Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2014)*, Pisa University Press, 2014, pp. 9–14.

[7] E. Bach, The algebra of events, *Linguistics and Philosophy* **9** (1986), 5–16.

[8] M. Ballesteros and J. Nivre, MaltOptimizer: An Optimization Tool for MaltParser, *in Proceedings of the System Demonstration Session of the 30th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012), ACL*, 2012, pp. 58–62.

[9] V. Bartalesi Lenzi, G. Moretti and R. Sprugnoli, CAT: The CELCT Annotation Tool, *in Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12), ELRA*, 2012, pp. 333–338.

[10] V. Basile, A. Bolioli, M. Nissim, V. Patti and P. Rosso, Overview of the EVALITA 2014 SENTIment POLarity Classification Task, *in Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2014)*, Pisa University Press, 2014, pp. 50–57.

[11] V. Basile and M. Nissim, Sentiment analysis on Italian tweets, *in Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, ACL*, 2013, pp. 100–107.

[12] S. Bethard, A Synchronous Context Free Grammar for Time Normalization, *in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, ACL*, 2013, pp. 821–826.

[13] B. Liu, *Sentiment analysis and subjectivity*, Taylor and Francis Group, Boca, 2010.

[14] J. Bollen and H. Mao, Twitter mood as a stock market predictor, *Computer* **44**(10) (2011), 91–94.

[15] B. Bohnet, Very High Accuracy and Fast Dependency Parsing is not a Contradiction, *in Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), ACL*, 2010, pp. 89–97.

[16] C. Bosco and A. Mazzei, The Evalita dependency parsing task from 2007 to 2011, *in: Evaluation of Natural Language and Speech Tools for Italian*, B. Magnini, F. Cutugno, M. Falcone and E. Pianta, eds., Springer–Verlag, Berlin Heidelberg, 2012, pp. 1–12.

[17] C. Bosco, S. Montemagni and M. Simi, Harmonization and merging of two Italian dependency treebanks, *in Proceedings of the LREC Workshop on Language Resource Merging, ELRA*, 2012, pp. 23–30.

[18] C. Bosco, S. Montemagni and M. Simi, Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank, *in Proceedings of the 7th ACL Linguistic Annotation Workshop and Interoperability with Discourse, ACL*, 2013, pp. 61–69.

[19] C. Bosco, V. Patti and A. Bolioli, Developing corpora for sentiment analysis: The case of irony and senti-TUT, *IEEE Intelligent Systems, Special Issue on Knowledge-based Approaches to Content-level Sentiment Analysis* **28**(2) (2013), 55–63.

[20] C. Bosco, L. Allisio, V. Mussa, V. Patti, G. Ruffo, M. Sanguinetti and E. Sulis, Detecting happiness in italian tweets: Towards an evaluation dataset for sentiment analysis in Felicittá, *in Proceedings of the 5th International Workshop on Emotion, Social Signals, Sentiment and Linked Open Data (ESSSLOD 2014), ELRA*, 2014, pp. 56–63.

[21] C. Bosco, P. Cosi, F. Dell'Orletta, M. Falcone, S. Montemagni and M. Simi eds., *Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2014)*, Pisa University Press, Pisa, 2014.

[22] S. Buchholz and E. Marsi, CoNLL-X Shared Task on Multilingual Dependency Parsing, *in Proceedings of the 10th Conference on Computational Natural Language Learning, ACL*, 2006, pp. 149–164.

[23] E. Cambria, B. Schuller, Y. Xia and C. Havasi, New avenues in opinion mining and sentiment analysis, *IEEE Intelligent Systems, Special Issue on Knowledge-based Approaches to Content-level Sentiment Analysis* **28**(2) (2013), 15–21.

[24] T. Caselli, *It-TimeML: TimeML annotation scheme for Italian, version 1.3.1*, Technical report, 2010.

[25] T. Caselli, V.B. Lenzi, R. Sprugnoli, E. Pianta and I. Prodanof, *Annotating events, temporal expressions and relations in Italian: The It-TimeML experience for the Ita-TimeBank, andin Proceedings of the 5th Linguistic Annotation Workshop (LAWV)*, 2011, pp. 143–151.

[26] T. Caselli, H. Llorens, B. Navarro-Colorado and E. Saquete, Data-driven approach using semantics for recognizing and classifying TimeML events in Italian, *in Proceedings of the International Conference on Recent Advances in Natural Language Processing 2011 (RANLP 2011), ACL*, 2011, pp. 533–538.

[27] T. Caselli, R. Sprugnoli, M. Speranza and M. Monachini, EVENTI. EValuation of Events and Temporal INformation at EVALITA 2014, *in Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2014)*, Pisa University Press, 2014, pp. 27–34.

[28] D. Chen and C. Manning, A Fast and Accurate Dependency Parser using Neural Networks, *in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), ACL*, 2014, pp. 740–750.

[29] N. Daniel, D. Radev and T. Allison, Sub-event based multi-document summarization, *in Proceedings of the HLT-NAACL'03 Workshop on Text summarization, workshop-Volume 5, ACL*, 2003, pp. 9–16.

[30] M.C. de Marneffe and C.D. Manning, Stanford Typed Dependencies representation, *in Proceedings of the COLING Work-shop on Cross-Framework and Cross-Domain Parser Evaluation (CrossParser 2008)*, ACL, 2008, pp. 1–8.

[31] M.C. de Marneffe and C.D. Manning, Stanford Typed Dependencies manual (Revised for the Stanford Parser v. 3.3 in December 2013), 2008, http://nlp.stanford.edu/software/dependencies_manual.pdf.

[32] M.C. DeMarneffe, T. Dozat, N. Silveira, K. Haverinen, F. Ginter, J. Nivre and C.D. Manning, Universal Stanford Dependencies: A Cross-Linguistic Typology, *in Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14), ELRA*, 2014, pp. 4585–4592.

[33] D. Davidov, O. Tsur and A. Rappoport, Semi-supervised recognition of sarcastic sentences in Twitter and Amazon, *in Proceedings of the 40th Conference on Computational Natural Language Learning (CoNLL 2010), ACL*, 2010, pp. 107–116.

[34] E. Diener, Subjective well-being: The science of happiness and a proposal for a national index, *American Psychologist* **55**(1) (2000), 34–43.

[35] X.L. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao and K. Murphy, Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion, *in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'14), ACM*, New York, 2014, pp. 601–610.

[36] J. Eisner, ed., *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007), ACL, Prague*, Czech Republic, 2007.

[37] A. Esuli, S. Baccianella and F. Sebastiani, Senti Word Net 3.0: An enhanced lexical resource for sentiment analysis and opinion mining, *in Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010), ELRA*, 2010, pp. 2200–2204.

[38] Y. Goldberg and M. Elhadad, An Efficient Algorithm for Easy-First Non-Directional Dependency Parsing, *in Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL (NAACL-2010), ACL*, 2010, pp. 742–750.

[39] J. Guldi and D. Armitage, *The History Manifesto*, Cambridge University Press, 2014.

[40] S. Husain, P. Mannem, B.R. Ambati and P. Gadde, The ICON-2010 tools contest on Indian language dependency parsing, *in Proceedings of ICON-2010 Tools Contest on Indian Language Dependency Parsing (ICON 2010)*, 2010, pp. 1–8.

[41] ISO, SemAf/Time Working Group, ISO DIS 24617-1:2008 Language resource management - Semantic annotation framework Part 1: Time and events, 2008, https://www.iso.org/obp/ui/#iso:std:iso:24617:-1:ed-1:v1:en.

[42] J. Katz-Brown, S. Petrov, R. McDonald, F. Och, D. Talbot, H. Ichikawa and M. Seno, Training a Parser for Machine Translation Reordering, *in Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP'11), ACL*, 2011, pp. 183–192.

[43] L. Kong and N.A. Smith, An Empirical Comparison of Parsing Methods for Stanford Dependencies, arXiv:1404.4314v1 [cs.CL], 2014.

[44] S. Kübler, The Page 2008 shared task on parsing German, *in Proceedings of the ACL-08 HLT Workshop on Parsing German (PaGe-08), ACL*, 2008, pp. 55–63.

[45] B. Magnini, F. Cutugno, M. Falcone and E. Pianta, eds., *Evaluation of Natural Language and Speech Tools for Italian*

-Proceedings of EVALITA 2011, LNCS/LNAI, Springer-Verlag, 2012.

[46] A.Martins, M. Almeida and N.A. Smith, Turning on the turbo: Fast third-order non–projective turbo parsers, *in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Volume 2: Short Papers, ACL*, 2013, pp. 617–622.

[47] D. Maynard and M. Greenwood, Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis, *in Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), ELRA*, 2014, pp. 4238–4243.

[48] R.T. McDonald, J. Nivre, Y. Quirmbach-Brundage, Y. Goldberg, D. Das, K. Ganchev, K.B. Hall, S. Petrov, H. Zhang, O. Täckström, C. Bedini, N. Bertomeu Castelló and J. Lee, Universal Dependency Annotation for Multilingual Parsing, *in Proceedings of 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), ACL*, 2013, pp. 92–97.

[49] L. Mitchell, M.R. Frank, K.D. Harris, P.S. Dodds and C.M. Danforth, The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place, *PLoS ONE* **8**(5), (2013), Public Library of Science, e64417.

[50] P. Nakov, S. Rosenthal, Z. Kozareva, V. Stoyanov, A. Ritter and T. Wilson, Semeval-2013 task 2: Sentiment analysis in Twitter, *in Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), ACL*, 2013, pp. 312–320.

[51] J. Nivre, J. Hall and J. Nilsson, MaltParser: A data-driven parser-generator for dependency parsing, *in Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), ELRA*, 2006, pp. 2216–2219.

[52] S. Oepen, M. Kuhlmann, Y. Miyao, D. Zeman, D. Flickinger, J. Hajič, A. Ivanova and Y. Zhang, SemEval 2014 Task 8: Broad-Coverage Semantic Dependency Parsing, *in Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), ACL*, 2014, pp. 63–72.

[53] B. Pang and L. Lee, Opinion mining and sentiment analysis, *in Foundations and Trends in Information Retrieval* **2**(1-2) (2008), 1–135.

[54] S. Petrov and R. McDonald, Overview of the 2012 shared task on parsing the web, *in Notes of the 1st Workshop on Syntactic Analysis of Non-Canonical Language (SANCL)*, vol. 59, 2012.

[55] E. Pianta, L. Bentivogli and C. Girardi, MultiWordNet: Developing an aligned multilingual database, *in Proceedings of the 1rst International Conference on Global Word Net*, 2002, pp. 293–302.

[56] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos and S. Manandhar, Semeval-2014 task 4: Aspect based sentiment analysis, *in Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), ACL*, 2014, pp. 27–35.

[57] J. Pustejovsky, J. Castao, R. Ingria, R. Saurí, R. Gaizauskas, A. Setzer and G. Katz, TimeML: Robust Specification of Event and Temporal Expressions in Text, *in Proceedings of the 5th International Workshop on Computational Semantics (IWCS-5)*, 2003.

[58] J. Pustejovsky, P. Hanks, R. Saurí, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro and M. Lazo, The TIMEBANK Corpus, *in Proceedings of Corpus Linguistics 2003, UCREL*, 2003, pp. 647–656.

[59] J. Pustejovsky, J. Littman, R. Saurí and R. Knippen, Temporal and event information in natural language text, *Language Resources and Evaluation* **39**(2-3) (2005), 123–164.

[60] *Proceedings of EVALITA*, Special Issue of Intelligenza artificiale, IV:2, 2007.

[61] *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*, Reggio Emilia, Italy, 2009.

[62] D. Quercia, J. Crowcroft, J. Ellis and L. Capra, Tracking "gross community happiness" from tweets, *in Proceedings of the ACM Conference on Computer Supported Cooperative Work), ACM*, 2012, pp. 965–968.

[63] A. Reyes, P. Rosso and T. Veale, A multidimensional approach for detecting irony in Twitter, *Language Resources and Evaluation* **47**(1) (2013), 239–268.

[64] L. Robaldo, T. Caselli, I. Russo and M. Grella, From Italian Text to TimeML Document via Dependency Parsing, in Computational Linguistics and Intelligent Text Processing, A. Gelbukh, ed., Lecture Notes in Computer Science 6609, Springer Berlin/Heidelberg, 2011, pp. 177–187.

[65] S. Rosenthal, A. Ritter, P. Nakov and V. Stoyanov, Semeval-2014 task 9: Sentiment analysis in Twitter, *in Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), ACL*, 2014, pp. 73–80.

[66] R. Saurí and J. Pustejovsky, FactBank: A corpus annotated with event factuality, *Language Resources and Evaluation* **43**(3) (2009), 227–268.

[67] F. Sartorio, G. Satta and J. Nivre, A Transition-Based Dependency Parser Using a Dynamic Parsing Strategy, *in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), ACL*, 2013, pp. 135–144.

[68] R. Saurí, *Annotating Temporal Relations in Catalan and Spanish*, TimeML Annotation Guidelines, Technical report, 2010.

[69] G. Savova, S. Bethard, W. Styler, J. Martin, M. Palmer, J. Masanz and W. Ward, Towards temporal relation discovery from the clinical narrative, *in American Medical Informatics Association (AMIA) annual symposium Proceedings*, 2009, pp. 568–572.

[70] D. Seddah, R. Tsarfaty, S.S. Kübler, et al., Overview of the SPRML 2013 shared task: Cross-framework evaluation of parsing morphologically rich languages, *in Proceedings of the 4th Workshop on Statistical Parsing of Morphologically Rich Languages, ACL*, 2013, pp. 146–182.

[71] M. Simi, C. Bosco and S. Montemagni, Less is More? Towards a Reduced Inventory of Categories for Training a Parser for the Italian Stanford Dependencies, *in Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), ELRA*, 2014, pp. 83–90.

[72] C. Strapparava and R. Mihalcea, Semeval-2007 task 14: Affective text, *in Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval 2007), ACL*, 2007, pp. 70–74.

[73] J. Strötgen and M. Gertz, HeidelTime: High Quality Rule-Based Extraction and Normalization of Temporal Expressions, *in Proceedings of SemEval 2010, ACL*, 2010, pp. 321–324.

[74] J. Strötgen, A. Armiti, T. Van Canh, J. Zell and M. Gertz, Time for more languages: Temporal tagging of Arabic, Italian, Spanish, and Vietnamese, *ACM Transactions on Asian Language Information Processing (TALIP), ACM*, **13**(1) (2014), 1–21.

[75] M. Surdeanu, R. Johansson, A. Meyers, L. Márquez and J. Nivre, The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies, *in CoNLL 2008: Proceedings of the 12th Conference on Computational Natural Language Learning*, 2008, pp. 159–177.

[76] A. Tumasjan, T.O. Sprenger, P.G. Sandner and I.M. Welpe, Predicting elections with Twitter: What 140 characters reveal about political sentiment, *in Proceedings of the 5th International*

AAAI Conference on Weblogs and Social Media (ICWSM-11), 2011, pp. 178–185.

[77] N. UzZaman, H. Llorens, L. Derczynski, J. Allen, M. Verhagen and J. Pustejovsky, Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations, *in 2nd Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), ACL*, 2013, pp. 1–9.

[78] S. Verma, S. Vieweg, W. Corvey, L. Palen, J.H. Martin, M. Palmer, A. Schram and K.M. Anderson, Natural language processing to the rescue? extracting "situational awareness" tweets during mass emergency, *in Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 2011, pp. 385–392.

[79] A. Volokh and G. Neumann, Task–oriented dependency parsing evaluation methodology, *in Proceedings of the IEEE 13th International Conference on Information Reuse & Integration (IRI)*, 2012, pp. 132–137.

[80] J. Wiebe, T. Wilson and C. Cardie, Annotating expressions of opinions and emotions in language, *Language Resources and Evaluation* **39**(2-3) (2005), 165–210.

[81] Y. Zhang and J. Nivre, Transition-based Dependency Parsing with Rich Non-local Features, *in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011), ACL*, 2011, pp. 188–193.