10.12871/clicit201425

EVENTI
EValuation of Events and Temporal INformation at Evalita 2014

Tommaso Caselli[*]
VU Amsterdam
De Boelelaan 1105, Amsterdam
t.caselli@gmail.com

Rachele Sprugnoli
FBK - University of Trento
Via Sommarive 18, Trento
sprugnoli@fbk.eu

Manuela Speranza
FBK
Via Sommarive 18, Trento
manspera@fbk.eu

Monica Monachini
ILC-CNR
Via G. Moruzzi 1, Pisa
monica.monachini@ilc.cnr.it

Abstract

English. This report describes the EVENTI (*EValuation of Events aNd Temporal Informatio*n) task organized within the EVALITA 2014 evaluation campaign. The EVENTI task aims at evaluating the performance of Temporal Information Processing systems on a corpus of Italian news articles. Motivations for the task, datasets, evaluation metrics, and results obtained by participating systems are presented and discussed.

Italiano. *Questo report descrive il task EVENTI (EValuation of Events aNd Temporal Information) organizzato nell'ambito della campagna di valutazione EVALITA 2014. EVENTI mira a valutare le prestazioni dei sistemi di processamento automatico dell'informazione temporale su un corpus di articoli di giornale in lingua italiana. Le motivazioni alla base del task, i dataset, le metriche di valutazione ed i risultati ottenuti dai sistemi partecipanti sono presentati e discussi.*

1 Introduction

Temporal Processing has recently become an active area of research in the NLP community. Reference to time is a pervasive phenomenon of human communication, and it is reflected in natural language. Newspaper articles, narratives and other text documents focus on events, their location in time, and their order of occurrence. Text comprehension itself involves, in large part, the ability to identify the events described in a text, locate them in time (and space), and relate them according to their order of occurrence. The ultimate goal of a temporal processing system is to identify all temporal elements (events, temporal expressions and temporal relations) either in a single document or across documents and provide a chronologically ordered representation of this information. Most NLP applications, such as Summarization, Question Answering, and Machine Translation, will benefit from such a capability. The TimeML Annotation Scheme (Pustejovsky et al., 2003a) and the release of annotated data have facilitated the development of temporally aware NLP tools. Similarly to what has been done in other areas of NLP, five open evaluation challenges[1] have been organized in the area of Temporal Processing. TempEval-2 has also boosted multilingual research in Temporal Processing by making TimeML compliant data sets available in six languages, including Italian. Unfortunately, partly due to the limited size (less than 30,000 tokens), no system was developed for Italian. Before the EVENTI challenge, there was no complete system for Temporal Processing in Italian, but only independent modules for event (Robaldo et al., 2011; Caselli et al., 2011b) and temporal expressions processing (HeidelTime) (Strötgen et al., 2014).

The EVENTI evaluation exercise[2] builds upon

[*] Formerly at Trento RISE

[1] TempEval-1: http://www.timeml.org/tempeval/; TempEval-2 http://timeml.org/tempeval2/; TempEval-3 http://www.cs.york.ac.uk/semeval-2013/task1/; TimeLine http://alt.qcri.org/semeval2015/task4/, and QA TempEval http://alt.qcri.org/semeval2015/task5/

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

previous evaluation campaigns to promote research in Temporal Processing for Italian by offering a complete set of tasks for comprehension of temporal information in written text. The exercise consists of a Main task on contemporary news and a Pilot task on historical texts and is based on the EVENTI corpus, which contains 3 datasets: the Main task training data, the Main task test data and the Pilot task test data.

2 EVENTI Annotation

The EVENTI exercise is based on the EVENTI annotation guidelines, a simplified version of the Italian TimeML Annotation Guidelines (henceforth, It-TimeML) (Caselli, 2010), using four It-TimeML tags: TIMEX3, EVENT, SIGNAL and TLINK. For clarity's sake, we report only the changes which have been applied to It-TimeML.

The TIMEX3 tag is used for the annotation of temporal expressions. No changes have been made with respect to It-TimeML.

The EVENT tag is used to annotate all mentions of events including verbs, nouns, prepositional phrases and adjectives. Changes concern the event extent. In particular, we have introduced exceptions to the minimal chunk rule for multi-token event expressions (the list of multi-token expressions created for this purpose is available online[3]). We have simplified the annotation of events realized by adjectives and prepositional phrases by restricting it to the cases in which they occur in predicate position with the explicit presence of a copula or a copular verb.

The SIGNAL tag identifies textual items which encode a relation either between EVENTs, or TIMEX3s or both. In EVENTI, we have annotated only SIGNALs indicating temporal relations.

The TLINK tag did not undergo any changes in terms of use and attribute values. Major changes concern the definition of the set of temporal elements that can be involved in a temporal relation. Details on this aspect are reported in the description of subtask C in Section 3.

3 EVENTI Subtasks

The EVENTI evaluation exercise is composed of a Main Task and a Pilot Task. Each task consists of a set of subtasks in line with previous TempEval campaigns and their annotation methodology.

The subtasks proposed are:

- Subtask A: determine the extent, the type and the value of temporal expressions (i.e. timex) in a text according to the It-TimeML TIMEX3 tag definition. For the first time, empty TIMEX3 tags were taken into account in the evaluation;

- Subtask B: determine the extent and the class of the events in a text according to the It-TimeML EVENT tag definition;

- Subtask C: identify temporal relations in raw text. This subtask involves 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. Given that EVENTI is an initial evaluation exercise in Italian and to avoid the difficulties of full temporal processing, we have further restricted this subtask by limiting the set of candidate pairs to: i.) pairs of main events in the same sentence; ii.) pairs of main event and subordinate event in the same sentence; and iii.) event - timex pairs in the same sentence. All temporal relation values in It-TimeML are used; i.e. BEFORE, AFTER, IS_INCLUDED, INCLUDES, SIMULTANEOUS, I(MMEDIATELY)_AFTER, I(MMEDIATELY)_BEFORE, IDENTITY, MEASURE, BEGINS, ENDS, BEGUN_BY and ENDED_BY.

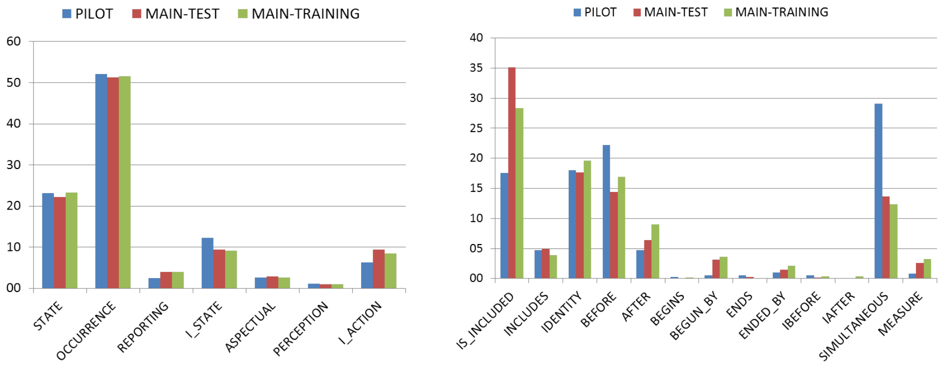
- Subtask D: determine the value of the temporal relation given two gold temporal elements (i.e. the source and the target of the relation) as defined in Task C (main event - main event; main event - subordinate event; event - timex).

4 Data Preparation and Distribution

The EVENTI evaluation exercise is based on the EVENTI corpus, which consists of 3 datasets: the Main task training data, the Main task test data and the Pilot task test data.

The news stories distributed for the Main task are taken from the Ita-TimeBank (Caselli et al., 2011a). Two expert annotators have conducted a manual revision of the annotations for the Main

[3] https://sites.google.com/site/eventievalita2014/data-tools/poliremEVENTI.txt



(a) *Event Class Values.*

(b) *Temporal Relations Values.*

Figure 1: Distribution of event classes and temporal relations in the EVENTI corpus (in percent).

task to solve inconsistencies mainly focusing on harmonizing event class and temporal relation values. The annotation revision has been performed using CAT[4] (Bartalesi Lenzi et al., 2012), a general-purpose web-based text annotation tool that provides an XML-based stand-off format as output. The final size of the EVENTI corpus for the Main task is 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[5] through the Meta-Share platform[6]. Annotated data are available under the Creative Commons Licence Attribution-NonCommercial-ShareAlike 3.0 to facilitate re-use and distribution for research purposes.

The Pilot test data consist of about 5,000 tokens from newspaper articles published in "*Il Trentino*" by Alcide De Gasperi, one of the founders of the Italian Republic and one of the fathers of the European Union (De Gasperi, 2006). All the selected news stories date back to 1914, the year of the outbreak of World War 1, a topic particularly relevant in 2014, the 100th anniversary of the Great War. They have been manually annotated in CAT by an expert annotator who followed the EVENTI Annotation Guidelines. As the aim of the Pilot task was to analyze how well systems built for contemporary languages perform on historical texts, no training data have been provided and participants were asked to participate with the systems developed for the Main task.

	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

Table 1: Annotated events, temporal expressions, signals and temporal relations in the EVENTI corpus.

Table 1 reports the total number of each annotated element type in the Main task training set, in the Main task test set, and in the Pilot test set.

	Main Training	Main Test	Pilot Test
EVENTs	172.1	142.4	239
TIMEX3s	26.4	23.3	19.0
TLINKs	33.7	39.7	76.4

Table 2: Average number of annotated events, temporal expressions and temporal relations per 1,000 tokens in the EVENTI corpus.

Table 2 presents the comparison between the average number of EVENTs, TIMEX3s and TLINKs annotated in the three datasets. The Pilot corpus clearly shows a higher density of events (238 vs. 172.1 and 142.4 for training and test, respectively) and temporal relations (76.4 vs. 33.7 and 39.7 for training and test, respectively). On the other hand, the average number of temporal expressions in the two corpora is comparable.

We illustrate in Figure 1 the distribution of the class values of EVENTs and the distribution of the temporal values for TLINKs. We can observe an even distribution of all classes among the three datasets. The most frequent classes are OCCURRENCE and STATE, followed by I_STATE and I_ACTION. The high prevalence of occurrences

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

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

[6] `http://www.meta-share.eu/`

and states is not surprising as these classes encode the objects of a narrative (e.g. contemporary news or historical texts) or what people "speak about". On the other hand, more interesting results are provided by the relatively high presence of the I_STATE and I_ACTION classes. According to the TimeML definitions, these classes are used either to express intensional relations or speculations about "possible worlds" between events. They are markers of subjectivity along the axis of event factivity, pointing out that people do not limit themselves to "speak about" happenings but they also speculate on these happenings. The higher frequency of I_STATE in the Pilot corpus with respect to the Main datasets is due to the fact that the Pilot dataset is mainly composed of editorial comments which frequently contain perspectives on and speculations about the world by the writer. Additional evidence is also the lower frequency of the REPORTING class in the Pilot dataset than in the Main task. The high presence of personal opinions influences also the temporal structure of the texts whereby most events are not ordered chronologically but presented as belonging to the same time frame on top of which the author expresses his opinions and suggests future and alternative courses of events. As a matter of fact, the most frequent temporal relation in the Pilot task is SIMULTANEOUS. On the other hand, in the Main task there is an evident preference for IS_INCLUDED. The main task is composed of news articles where events tend to be more often linked to temporal containers (e.g. temporal expressions or other events) to facilitate understanding of stories by readers.

5 Evaluation

Given the strong connection of this task with the TempEval Evaluation Exercises, we adopted the evaluation metrics developed in TempEval-3 (UzZaman et al., 2013) with minor modifications[7]. In particular, the scorer was adapted in order to take CAT files as input and the evaluation of temporal expressions was extended to include empty TIMEX3 tags.

Concerning the temporal elements in subtask A and subtask B, we evaluated: i) the number of the elements correctly identified and if their extension is correct, and ii.) the attribute values correctly identified. For recognition, we used Precision, Recall and F1-score. Strict and relaxed match were both taken into account. As for attribute evaluation, we used F1-score to measure how well a system identifies an element and its attribute values. For subtask A, we computed Attribute F1-score on VALUE and Attribute F1-score on TYPE, and based the final ranking on the former. For subtask B, we computed attribute F1-score on CLASS, on which we based the final ranking.

For subtask C, we took into consideration three aspects : i) the number and the extent of the temporal elements identified in a raw text ii) the identification of the correct sources and targets applying both strict and relaxed match and iii) the identification of the correct temporal value. In subtask D, we evaluated only the identification of the correct temporal value. Similarly to subtasks A and B, we computed Precision, Recall and F1-score also for subtasks C and D and we set the final rankings on the basis of F-1 scores[8].

6 Participant Systems

Although eight teams registered for the task, only three actually submitted the output of their systems for a total of 17 unique runs: FBK (Fondazione Bruno Kessler), HT (University of Heidelberg), and UNIPI (Università di Pisa). We report below a short description of the systems the three teams developed. Detailed descriptions are reported in the system papers of the Evalita 2014 Proceedings (Bosco et al., 2014).

FBK is an end-to-end system based on a machine learning approach, namely supervised classification. It was developed for the EVENTI exercise by combining and adapting to Italian three subsystems first developed for English within the NewsReader project[9]: one for time expression recognition and normalization, one for event extraction, and one for temporal relation identification and classification. Temporal expression recognition and classification is conducted by means of an adaptation to Italian of TimeNorm (Bethard, 2013), a rule-based system based on synchronous context free grammars. The other subsystems are based on machine learning and use a Support Vector Machine approach.

HeidelTime is a rule-based, multilingual and

[7] The scorer of EVENTI is available online: `http://goo.gl/TbnE7D`

[8] TLINK directionality was not an issue as the scorer is able to deal with reciprocal temporal relations

[9] `http://www.newsreader-project.eu`

		RECOGNITION				NORMALIZATION	
		F1	P	R	Strict F1	TYPE F1	VALUE F1
MAIN TASK	HT 1.7	0.78	0.921	0.676	0.662	0.643	0.571
	HT 1.8	0.893	0.935	0.854	0.821	0.643	**0.709**
	HT 1.8 (no ET)	0.878	0.94	0.824	0.804	0.775	0.69
	FBK_A1	0.886	0.936	0.841	0.827	0.8	0.665
	UNIPI_1	0.768	0.929	0.654	0.662	0.643	0.566
	UNIPI_2	0.771	0.922	0.662	0.659	0.64	0.563
PILOT TASK	HT 1.7	0.653	0.96	0.495	0.585	0.571	0.408
	HT 1.8	0.788	0.918	0.691	0.671	0.624	0.459
	HT 1.8 (no ET)	0.781	0.917	0.68	0.663	0.615	0.45
	FBK_A1	0.87	0.963	0.794	0.746	0.678	**0.475**

Table 3: Results of Main and Pilot tasks for subtask A - TIMEX3s recognition and normalization.

		RECOGNITION				CLASS
		F1	P	R	Strict F1	F1
MAIN TASK	FBK_B1	0.884	0.902	0.868	0.867	**0.671**
	FBK_B2	0.749	0.917	0.632	0.732	0.632
	FBK_B3	0.875	0.915	0.838	0.858	0.67
PILOT TASK	FBK_B1	0.843	0.9	0.793	0.834	**0.604**
	FBK_B2	0.681	0.897	0.548	0.671	0.535
	FBK_B3	0.83	0.92	0.756	0.819	0.602

Table 4: Results of Main and Pilot tasks for subtask B - Events recognition and *class* assignment.

		F1	P	R	Strict F1
MAIN TASK	FBK_C1 (B1_D1)	**0.264**	0.296	0.238	0.341
	FBK_C2 (B1_D2)	0.253	0.265	0.241	0.325
	FBK_C3 (B2_D1)	0.209	0.282	0.167	0.267
	FBK_C4 (B2_D2)	0.168	0.203	0.255	0.258
	FBK_C5 (B3_D1)	0.247	0.297	0.211	0.327
	FBK_C6 (B3_D2)	0.247	0.297	0.211	0.327
PILOT TASK	FBK_C1 (B1_D1)	**0.185**	0.277	0.139	0.232
	FBK_C2 (B1_D2)	0.174	0.233	0.139	0.221
	FBK_C3 (B2_D1)	0.141	0.243	0.099	0.178
	FBK_C4 (B2_D2)	0.139	0.215	0.102	0.174
	FBK_C5 (B3_D1)	0.164	0.268	0.118	0.209
	FBK_C6 (B3_D2)	0.164	0.268	0.118	0.209

Table 5: Results of Main and Pilot tasks for subtask C - Temporal relations from raw texts.

cross-domain temporal tagger initially developed for English in the context of TempEval-2 (Strötgen and Gertz, 2010), which makes use of regular expressions. The distributed version of HeidelTime, which is freely available under a GNU General Public License, already supports Italian temporal tagging. For the EVENTI exercise, HT extended HeidelTime by tackling the recognition of TimeML's empty TIMEX3 tags and by tuning HeidelTime's Italian resources (e.g. by extending patterns, adding rules, and improving existing ones) on the basis of the more specific annotation guidelines and the training data released by the task organizers.

UNIPI used the available version of HeidelTime and adapted it by integrating into the pipeline the Tanl tools (Attardi et al., 2010), a suite of statistical machine learning tools for text analytics based on the software architecture paradigm of data pipelines.

7 System Results

For subtask A, temporal expression recognition and normalization, we had 3 participants and 6 unique runs. Table 3 shows the results for both the Main and the Pilot tasks. In the Main Task, only the best scoring run, i.e. HT 1.8, achieved results in terms of F1 above 0.70 in the normalization of the VALUE attribute. However, in the assignment of the TYPE attribute, FBK_A1 outperformed it (0.8 vs. 0.643). As for recognition, all the runs have a precision above 0.92, while recall ranges from 0.654 to 0.854. An analogous trend in the recognition of temporal expressions was registered in the Pilot task. The best run proved to be FBK_A1 with a VALUE F1 of 0.475.

Only one team participated in the remaining three subtasks. In subtask B, event detection and classification, 3 different runs were submitted. The evaluation results are reported in Table 4. FBK_B1 is the best run both in the Main task and in the Pilot task with an F1 on class assignment of 0.671 and 0.604 respectively. FBK_B1 has the best results also in terms of event recognition (0.884 in the Main task and 0.843 in the Pilot task). Precision in event recognition is high, above 0.89, in both tasks. Recall, on the other hand, ranges from 0.548, the lowest score obtained in the Pilot task, to 0.868, the highest score obtained in the Main task.

Results of Main and Pilot tasks for subtask C, i.e. temporal relations from raw texts, are reported in Table 5. For both Main task and Pilot task, the best performing run is FBK_C1, with 0.264 F-score and 0.185 F-score respectively.

In subtask D, i.e. TLINKs with temporal elements given, two runs were submitted. As shown in Table 6, FBK_D1 performed better than FBK_D2 with a difference of more than 0.3 points (0.736 vs. 0.419).

	F1	P	R	Strict F1
FBK_D1	**0.736**	0.74	0.731	0.731
FBK_D2	0.419	0.342	0.541	0.309

Table 6: Results of Main and Pilot tasks for subtask D - TLINKs with temporal elements given.

8 Discussion

EVENTI achieved a significant result in setting the state of the art on Temporal Processing for Italian although the reduced number of participants for three of the four subtasks limits observations on the participants' results.

Subtask A, temporal expression recognition and normalization, attracted the highest number of participants. Two participants, HT and UNIPI, developed rule-based systems both for recognition and normalization and submitted three and two runs respectively: HT 1.7 (the HT system publicly available), HT 1.8 (the system adapted to EVENTI) , HT 1.8 (the adapted system wothout the empty tag feature), UNIPI_1 (a baseline obtained by using the same publicly available system as HT 1.7), and UNIPI_2 (obtained substituting the TreeTagger with the Tanl Tokenizer in HeidelTime). FBK, on the other hand, developed a hybrid system: recognition is conducted by means of an SVM classifier while normalization is provided by a rule based system adapted to Italian (TimeNorm). Concerning recognition of temporal expressions, competition among the best performinig systems, HT 1.8 and FBK_A1, is high (the difference in performance is less than 1%). On the Main task data (contemporary news articles), the statistical system, FBK_A1, performs best at strict matching, and only one rule-based system, HT 1.8, performs best at relaxed matching. The difference in performance between the two rule based systems, HT and UNIPI_2, both for recognition and normalization clearly points to a problem in the integration of the Tanl POS tagset in the HT system, rather than signaling a limit of the approach for this task. Unfortunately, it is not possible to compare these results with those obtained by the systems participating in the EVALITA 2007 TERN (*Temporal Expression Recognition and Normalization*) Task (Bartalesi Lenzi and Sprugnoli, 2007) for two main reasons: firstly, the annotation of TIMEX3 tags substantially differs from that for TIMEX2, which was used for TERN, in terms of tag spans, normalization and presence of empty timex tags; and secondly, the evaluation methods in TERN, except for the recognition task, are not comparable with those used in EVENTI.

Subtask B, event detection and classification, had only one team with 3 different runs. The FBK system is based on an SVM classifier. The difference in performance between the three runs does not concern the features used for training but the classification method. The best result, FBK_B1's strict F1 0.867, was obtained by splitting the detection and classification task into two steps, first detection and then classification, and using a one-vs-one strategy. In the classification task, the predictions of the detection classifier were incorporated as a feature. FBK_B3, which obtained comparable results to FBK_B1, implements a single classifier with one-vs-rest multi-class classification. Difference in performance is less than 1% suggesting that both approaches are highly competitive but require different multi-class classification methods. Semantics is encoded by means of lexical knowledge through MultiWordNet (Pianta et al., 2002). Comparisons with (Caselli et al., 2011b) and (Robaldo et al., 2011) are not possible due to the different sizes of the training and test sets and also because the original TempEval-2 test set for Italian has been incorporated in the EVENTI training set. Nevertheless, the results reported in (Caselli et al., 2011b) for event classes suggest that more fine grained and specialized lexical knowledge for event classification may provide better results.

Subtasks C and D are focused on temporal relations. The unique participant, i.e. FBK, submitted 6 runs for subtask C and 2 for subtask D. The system for subtask C tackles the task in a two step approach: first an SVM classifier identifies all eligible event-event and event-timex pairs for a temporal relation. Subsequently, a second SVM classifier, based on a previous framework for temporal relations between entities (Mirza and Tonelli, 2014), assigns the temporal relations values. This classifier mostly uses basic morphosyntactic features plus additional information based on the annotated SIGNAL. Different versions of the system (FBK_C2, FBK_C4, FBK_C6 and FBK_D2) incorporate TLINK rules for event-timex pairs which include signals as reported in the annotation guidelines. The system for subtask D corresponds to the second SVM classifier developed for subtask C. In both subtasks the presence of rules for event-timex temporal relations have a negative impact on system performance.

Concerning the Pilot task, no comparisons with previous evaluations can be drawn. To the best of our knowledge, EVENTI is the first evaluation exercise on Temporal Information Processing on historical texts. In general, a drop in the systems' performance was registered. In particular, the drop in the normalization of temporal expressions can probably be explained by the fact that 54% of the temporal expressions in the Pilot corpus is fuzzy (e.g. *i sacrifici dell'*⟨*ora presente*⟩) or non-specific (e.g. *nei* ⟨*giorni*⟩ *del dolore*), with respect to 24% in the Ita-TimeBank. A similar decrease in performance was registered in subtask D, submitted post evaluation by FBK, where both runs achieved an F1-score of 0.57.

8.1 Comparison with TempEval-3

Although no direct comparison can be made, it is still interesting to compare the performance among systems in different languages, developed and tested on annotation schemes which are compliant with a common standard (i.e. ISO-TimeML). We report in Table 7 the results of the best systems from TempEval-3 (UzZaman et al., 2013) for English (EN) and Spanish (ES) with respect to the identification of temporal relation from raw text.

		Strict F1	F1 attribute
TASK A	HT 1.8	0.893	0.709
	HeidelTime_EN	0.813	0.776
	HeidelTime_ES	0.853	0.875
TASK B	FBK_B1	0.867	0.671
	ATT-1_EN	0.810	0.718
	TIPSemB-F_ES	0.888	0.576
TASK C∗	FBK_C1	0.341	0.264
	ClearTK-2_EN	*n.a.*	0.309
	TIPSemB-F_ES	*n.a.*	0.416
TASK D∗	FBK_D1	0.731	0.736
	UTTime-1, 4_EN	*n.a.*	0.564

Table 7: Comparison with TempEval-3 systems.

Results for temporal expression detection, Task A, are above 0.80 in all languages. The results for normalization present a higher variability ranging from 0.709 for Italian up to 0.875 for Spanish. The lower results for Italian can be due to the fact that empty TIMEX3 tags were taken into account in the evaluation, while this was not done in TempEval-3. Still the difference between English and Italian is minor when compared to Spanish.

In Task B, event detection and normalization, system results are pretty similar for event detection but differ highly for the classification. This difference can be due mainly to the annotated data as all systems are comparable in terms of features used.

Finally, the analysis of Task D and C requires a *caveat*, namely that Task C, full temporal processing, has been simplified in Italian with respect to Task C in TempEval-3. Nevertheless, the results are very low, signaling that this task is very hard and that different approaches and solutions are to be envisaged.

9 Conclusion

This paper describes the EVENTI evaluation exercise within the EVALITA 2014 evaluation campaign. The task requires the participants to automatically annotate a raw text with temporal information. This involves the identification of temporal expressions, events and temporal relations. As for temporal relations, we have restricted the set of relations only to event-event and event-timex pairs in the same sentence.

The EVENTI evaluation exercise is the first end-to-end task on Temporal Processing for Italian and it is strictly linked to the TempEval-3 challenge. In particular, it adopts the same evaluation method thus aiming at facilitating comparison between systems developed in different languages. EVENTI is also the first evaluation on Temporal Processing of Historical Texts, organized to foster the collaboration between the NLP and the Digital Humanities communities.

Future work will aim at providing the full set of temporal relations without restrictions and possibly investigate temporal processing in specific applications or broader tasks (e.g. RTE and QA) both for Italian and from a multilingual perspective. The results obtained by the one end-to-end system participating in EVENTI show that there is still room for improvement in the identification and interpretation of temporal expressions, events, and temporal relations.

10 Acknowledgments

Our thanks to Nashaud UzZaman which has allowed us to re-use the evaluation script of TempEval-3 for the EVENTI Task, Giovanni Moretti for his assistence in trasforming the data to the CAT format, Anne-Lyse Minard for adapting the evaluation script.

References

G. Attardi, S. Dei Rossi, and M. Simi. 2010. The Tanl Pipeline. In *Proc. of LREC Workshop on WSPP*.

V. Bartalesi Lenzi and R. Sprugnoli. 2007. Evalita 2007: Description and Results of the TERN Task. *Intelligenza artificiale*, 2(IV):55–57.

V. Bartalesi Lenzi, G. Moretti, and R. Sprugnoli. 2012. CAT: the CELCT Annotation Tool. In *Proceedings of the Eighth International conference on Language Resources and Evaluation (LREC-12)*, pages 333–338.

S. Bethard. 2013. A Synchronous Context Free Grammar for Time Normalization. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 821–826, Seattle, Washington, USA, October. Association for Computational Linguistics.

C. Bosco, F. DellOrletta, S. Montemagni, and M. Simi, editors. 2014. *Evaluation of Natural Language and Speech Tools for Italian*, volume 1. Pisa University Press.

T. Caselli, V.B. Lenzi, R. Sprugnoli, E. Pianta, and I. Prodanof. 2011a. Annotating events, temporal expressions and relations in italian: the it-TimeML experience for the Ita-TimeBank. In *Proceedings of the 5th Linguistic Annotation Workshop (LAW V)*, pages 143–151.

T. Caselli, H. Llorens, B. Navarro-Colorado, and E Saquete. 2011b. Data-driven approach using semantics for recognizing and classifying TimeML events in Italian. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 533–538.

T. Caselli. 2010. IT-TimeML: TimeML annotation scheme for Italian, version 1.3.1, technical report. Technical report, ILC-CNR, Pisa.

A. De Gasperi. 2006. Scritti e discorsi politici. In E. Tonezzer, M. Bigaran, and M. Guiotto, editors, *Scritti e discorsi politici*, volume 1. Il Mulino.

P. Mirza and S. Tonelli. 2014. Classifying Temporal Relations with Simple Features. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden.

E. Pianta, L. Bentivogli, and C. Girardi. 2002. MultiWordNet: developing an aligned multilingual database. In *First International Conference on Global WordNet*, Mysore, India.

J. Pustejovsky, J. Castao, R. Ingria, R. Saurì, R. Gaizauskas, A. Setzer, and G. Katz. 2003a. TimeML: Robust Specification of Event and Temporal Expressions in Text. In *Fifth International Workshop on Computational Semantics (IWCS-5)*.

J. Pustejovsky, P. Hanks, R. Saurí, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, and M. Lazo. 2003b. The TIMEBANK corpus. In *Corpus Linguistics 2003*.

L. Robaldo, T. Caselli, I. Russo, and M. Grella. 2011. From Italian Text to TimeML Document via Dependency Parsing. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 6609 of *Lecture Notes in Computer Science*, pages 177–187. Springer Berlin / Heidelberg.

J. Strötgen and M. Gertz. 2010. HeidelTime: High Quality Rule-Based Extraction and Normalization of Temporal Expressions. In *Proceedings of SemEval 2010*, pages 321–324, Uppsala, Sweden, July. Association for Computational Linguistics.

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

N. UzZaman, H. Llorens, L. Derczynski, J. Allen, M. Verhagen, and J. Pustejovsky. 2013. SemEval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Proceedings of SemEval-2013*, pages 1–9. Association for Computational Linguistics, Atlanta, Georgia, USA.