Evaluation tasks in open competitions

Deliverable D10.4

Version FINAL

Authors: Anne-Lyse Minard[2], Manuela Speranza[2], Marieke van Erp[1], Antske Fokkens[1], Marten Postma[1], Piek Vossen[1], Eneko Agirre[3], Itziar Aldabe[3], German Rigau[3], Ruben Urizar[3]

Affiliation: (1) VUA, (2) FBK, (3) EHU

Building structured event indexes of large volumes of financial and economic data for decision making

ICT 316404

Grant Agreement No.	316404
Project Acronym	NEWSREADER
Project Full Title	Building structured event indexes of large volumes of financial and economic data for decision making.
Funding Scheme	FP7-ICT-2011-8
Project Website	http://www.newsreader-project.eu/
Project Coordinator	Prof. dr. Piek T.J.M. Vossen VU University Amsterdam Tel. + 31 (0) 20 5986466 Fax. + 31 (0) 20 5986500 Email: piek.vossen@vu.nl
Document Number	Deliverable D10.4
Status & Version	FINAL
Contractual Date of Delivery	July 2015
Actual Date of Delivery	February 1, 2016
Type	Report
Security (distribution level)	Public
Number of Pages	34
WP Contributing to the Deliverable	WP10
WP Responsible	FBK
EC Project Officer	Susan Fraser
Authors: Anne-Lyse Minard[2], Manuela Speranza[2], Marieke van Erp[1], Antske Fokkens[1], Marten Postma[1], Piek Vossen[1], Eneko Agirre[3], Itziar Aldabe[3], German Rigau[3], Ruben Urizar[3]	
Keywords: TimeLines, Open competitions, SemEval, EVENTI, NewsStory, CLIN, temporal processing	
Abstract: This deliverable describes the evaluation tasks organized in the project. We present the SemEval task "TimeLine" that was organized by the NewsReader project, the workshop "NewsStory" at ACL, the Dutch Shared Task at CLIN26 and the EVENTI task for Italian endorsed by the project.	

Table of Revisions

Version	Date	Description and reason	By	Affected sections
0.0	21 May 2015	Init	Anne-Lyse Minard	all
0.1	11 June 2015	Added section about EVENTI	Manuela Speranza	4
0.2	22 June 2015	Added introduction and conclusion sections, improvement of the overall	Anne-Lyse Minard	all
1.0	30 June 2015	Internal review	Ruben Urizar	all
1.1	23 July 2015	Evalita 2016	Manuela Speranza and Anne-Lyse Minard	6
1.2	29 July 2015	CLIN26, references	Antske Fokkens	6, References
1.2	31 July 2015	Check by coordinator	VUA	-
2.0	14 January 2016	CLIN26 results	Antske Fokkens	6
2.1	19 January 2016	Internal review	Anne-Lyse Minard	all
2.1	29 January 2016	Check by coordinator	VUA	-

Executive Summary

This deliverable describes three evaluation tasks in open competitions organized during the second and third years of the NewsReader project.

At SemEval2015 (Semantic evaluation exercises), we organized the first task about timeline creation for English: "TimeLine: cross-document event ordering". The data needed for this task were prepared in the WP3. This task sets up a first framework for evaluating timeline creation systems. A workshop at ACL was organized and aimed to lead a discussion about the definition of a storyline, the annotation of storylines and the evaluation of storyline extraction taking as starting point the SemEval TimeLine task.

The NewsReader project also supported a task about temporal processing in Italian at the evaluation campaign Evalita 2014. At the end of the third year of the project, we organized the first Dutch Shared Task at CLIN26. The data annotated within NewsReader for Dutch was used as development and evaluation datasets. The evaluation exercise included tasks on Named Entity Recognition and Classification, Nominal coreference and Event recognition and factuality.

Finally, we plan to propose a task at Evalita 2016 on Italian using the data annotated within the WP3 as test data.

Contents

1 Introduction

The deliverable describes the evaluation tasks in open competitions that NewsReader's partners have organized or supported. The goal of the NewsReader project is to reconstruct event story lines from the news by automatically processing daily news streams. In this context, we organized a SemEval task on timelines creation in English: "TimeLine: Cross-Document Event Ordering", a workshop at ACL 2015 on computing news storylines: "the 1st workshop on Computing News Storylines" and the first CLIN Dutch Shared Task at CLIN26, and we have supporteed a task on temporal processing in Italian at Evalita 2014: "Evaluation of Events and Temporal Information".

This deliverable is structured as follow. In Section 2, we describe the SemEval task on timelines creation, followed by the paper describing the task published in the proceedings of the SemEval workshop. In Section 3, we present the workshop "Computing News Story-Lines" that was held during the ACL conference in 2015. The task on Temporal Processing in Italian endorsed by NewsReader is described in Section 4, which also contains the task description paper published in Evalita 2014 proceedings. In Section 5 we present the task organized at CLIN26 for Dutch. Finally in Section 6 we present the task to be proposed at Evalita 2016 for Italian.

2 SemEval Task: TimeLine

As part of going beyond document-based evaluations, the NewsReader team set up a Timeline evaluation in the context of the SemEval-2015: Semantic Evaluation Exercises.[1] The task, "TimeLine: Cross-Document Event Ordering" was accepted as a pilot task in order to gauge state-of-the-art cross-document timeline creation. Given a set of documents and a set of target entities, the task consisted in building a timeline for each entity, by detecting, anchoring in time and ordering the events involving that entity.

Before the evaluation, we made available to the participants the Task guidelines, the trial dataset and the evalution tool (first version on May 30, 2014). The trial dataset as well as the evaluation dataset had been annotated within the NewsReader project following the Task guidelines (see Deliverable D3.3.2 (van Erp *et al.* (2015)) for more information about the data).

The research community showed a great interest in TimeLine, with 27 teams (from 19 different institutions and 15 different countries) signing up for the evaluation task. Out of these, 8 actually downloaded the evaluation dataset and 4 submitted their systems' results, for a total of 13 unique runs. The teams that submitted their systems' results were the following: WHUNLP from Wuhan University in China, SPINOZAVU from VU University Amsterdam in the Netherdland, GPLSIUA from University of Alicante in Spain, and HEIDELTOUL from Heidelberg University in Germany.

In the main track (timeline creation from raw text), the best score was obtained by WHUNLP (with an F1-score of 7.28%). In Track B, for which the input texts contained

[1] `http://alt.qcri.org/semeval2015/`

event mention annotations, the best score was obtained by GPLSIUA (with an F1-score of 25.36%). In Subtracks A and B, which had been proposed to evaluate systems that do not perform time normalisation or event anchoring in time but focus on temporal relations between events, the best results were obtained respectively by SPINOZAVU – who achieved an F1-score of 1.69% – and by GPLSIUA – who achieved an F1-score of 23.15%. The complete results were presented to the 9th International Workshop on Semantic Evaluations (SemEval 2015) during the NAACL-HLT conference at Denver, Colorado on June 4 and 5, 2015. The team SPINOZAVU did an oral presentation of their system during the workshop and the team GPLSIUA presented its system through a poster.

The schedule of the task was the following:

- Trial data ready: May 30, 2014
- Evaluation start: December 10, 2014
- Evaluation end: December 17, 2014
- Paper submission due: January 30, 2015
- Paper reviews due: February 28, 2015
- Camera ready due: March 30, 2015
- SemEval workshop: June 4-5, 2015

The task was organized by Anne-Lyse Minard, Manuela Speranza, Eneko Agirre, Itziar Aldabe, Marieke van Erp, Bernardo Magnini, German Rigau and Ruben Urizar.

All the materials made available for the participants as well as the results can be accessed from the SemEval website: `http://alt.qcri.org/semeval2015/task4/`.

The full task description and annotation guidelines can be found at `http://www.newsreader-project.eu/publications/technical-reports/` in the following technical reports:

- Anne-Lyse Minard, Manuela Speranza, Bernardo Magnini, Marieke van Erp, Itziar Aldabe, Ruben Urizar, Eneko Agirre and German Rigau. *TimeLine: Cross-Document Event Ordering. SemEval 2015 – Task 4.* NWR-2014-10. Fondazione Bruno Kessler.
- Anne-Lyse Minard, Alessandro Marchetti, Manuela Speranza, Bernardo Magnini, Marieke van Erp, Itziar Aldabe, Ruben Urizar, Eneko Agirre and German Rigau. *TimeLine: Cross-Document Event Ordering. SemEval 2015 – Task 4. Annotation Guidelines.* NWR-2014-11. Fondazione Bruno Kessler.

In the remainder of the section we have included the paper published in the proceedings of the the 9th International Workshop on Semantic Evaluations (SemEval 2015).

SemEval-2015 Task 4: TimeLine: Cross-Document Event Ordering

Anne-Lyse Minard[1], **Manuela Speranza**[1], **Eneko Agirre**[2], **Itziar Aldabe**[2], **Marieke van Erp**[3], **Bernardo Magnini**[1], **German Rigau**[2], **Rubén Urizar**[2]
[1] Fondazione Bruno Kessler, Trento, Italy
[2] The University of the Basque Country (UPV/EHU), Spain
[3] VU University Amsterdam, the Netherlands
{minard,manspera,magnini}@fbk.eu, marieke.van.erp@vu.nl
{itziar.aldabe,e.agirre,german.rigau,ruben.urizar}@ehu.eus

Abstract

This paper describes the outcomes of the TimeLine task (Cross-Document Event Ordering), that was organised within the Time and Space track of SemEval-2015. Given a set of documents and a set of target entities, the task consisted of building a timeline for each entity, by detecting, anchoring in time and ordering the events involving that entity. The TimeLine task goes a step further than previous evaluation challenges by requiring participant systems to perform both event coreference and temporal relation extraction across documents. Four teams submitted the output of their systems to the four proposed subtracks for a total of 13 runs, the best of which obtained an F_1-score of 7.85 in the main track (timeline creation from raw text).

1 Introduction

In any domain, it is important that professionals have access to high quality knowledge for taking well-informed decisions. As daily tasks of information professionals revolve around reconstructing a chain of previous events, an insightful way of presenting information to them is by means of timelines. The aim of the Cross-Document Event Ordering task is to build timelines from English news articles. To provide focus to the timeline creation, the task is presented as an ordering task in which events involving a particular target entity are to be ordered chronologically. The task focuses on cross-document event coreference resolution and cross-document temporal relation extraction.

The latter has been the topic of the three previous TempEval tasks within the SemEval challenges:

- TempEval-1 (2007): Temporal Relation Identification (Verhagen et al., 2009)
- TempEval-2 (2010): Evaluating Events, Time Expressions, and Temporal Relations (Verhagen et al., 2010)
- TempEval-3 (2013): Temporal Annotation (UzZaman et al., 2013)

Additionally, it has also been the focus of the 6th i2b2 NLP Challenge for clinical records (Sun et al., 2013). The cross-document aspect, however, has not often been explored. One example is the work described in (Ji et al., 2009) using the ACE 2005 training corpora. Here the authors link pre-defined events involving the same centroid entities (i.e. entities frequently participating in events) on a timeline. Nominal coreference resolution has been the topic of SemEval 2010 Task on Coreference Resolution in Multiple Languages (Recasens et al., 2010). TimeLine is a pilot task that goes beyond the above-mentioned evaluation exercises by addressing coreference resolution for events and temporal relation extraction at a cross document level.

This task was motivated by work done in the NewsReader project[1]. The goal of the NewsReader project is to reconstruct story lines across news articles in order to provide policy and decision makers with an overview of what happened, to whom, when, and where. Thus, the NewsReader project aims to present end-users with cross-document storylines. Timelines are intermediate event represen-

[1] http://www.newsreader-project.eu

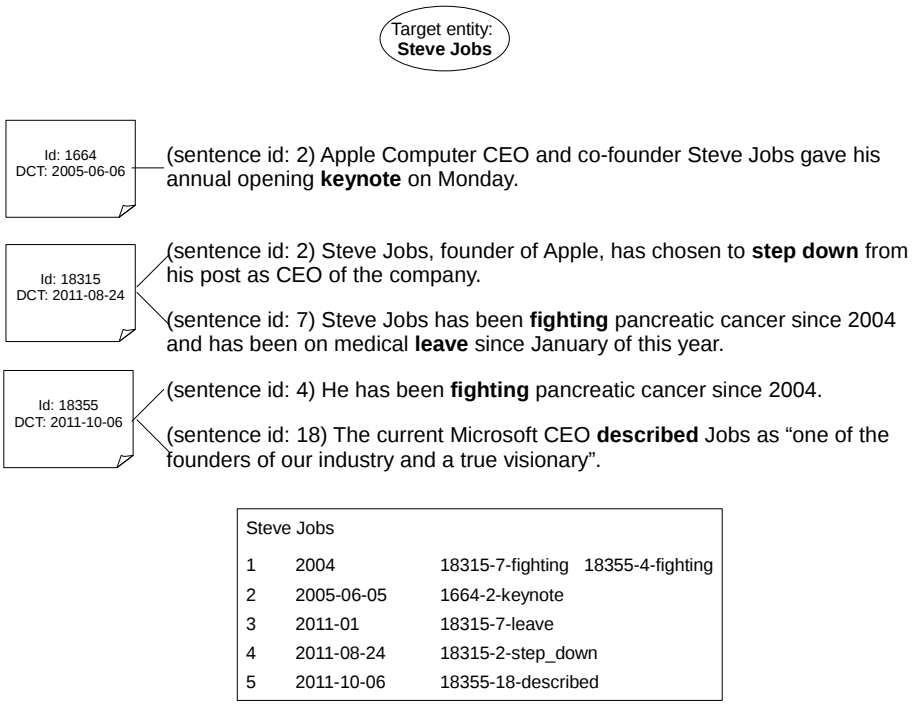


Figure 1: Example of a timeline for the target entity "Steve Jobs" built from five sentences coming from three documents.

tations towards this goal.

The remainder of this paper is organised as follows. In Section 2, we introduce the task. In Section 3, we describe the data annotation protocol. In Section 4, we present the characteristics of our dataset and gold standard timelines. In Section 5, we describe our evaluation methodology, followed by the description of participant systems in Section 6 and the results obtained by the participants to the task in Section 7. Lessons learnt and limitations of our setup are discussed in Section 8.

2 Task Description

Given a set of documents and a set of target entities, the task consists of building a timeline related to each entity, i.e. detecting, anchoring in time, and ordering the events in which the target entity is involved (Minard et al., 2014b). We base our notion of event on TimeML, according to which an *event* is a cover term for situations that happen or occur, including predicates describing states or circumstances in which something obtains or holds true (Pustejovsky et al., 2003).

As input data, we provide a set of documents and a set of target entities; only entities involved in more than two events across at least two different documents are considered as candidates target entities. We also propose two different tracks on the basis of the data used as input: **Track A**, for which we provided only the raw text sources (main track), and **Track B**, for which we also made gold event mentions available.

The expected output, both for Track A and B, is one timeline for each target entity. A timeline for a specific target entity consists of the ordered list of the events in which that entity participates. Events in a timeline are anchored in time through the time anchor attribute; however, for both Track A and B, we also propose a subtrack in which the events do not need to be associated to a time anchor.

In Figure 1 we show an example of a timeline for the target entity *Steve Jobs* built using five sentences extracted from three documents. In bold we represent the events that form the timeline.

In order to perform the task, participants are required to resolve entity coreference, as timelines should contain events involving all coreferring textual mentions of the target entities (including pronominal mentions). For example, in Figure 1, the event *fighting* involving the target entity *Steve Jobs* mentioned as *he* is included in the timeline together with other events also referring to *Steve Jobs*.

The dataset released for this task is composed of 120 Wikinews[2] articles and 44 target entities. 30 documents and 6 target entities (each associated to a timeline) are provided as trial data, while the evaluation dataset consist of 90 documents and 38 target entities (each associated to a timeline).

3 Data Annotation

We manually selected a set of target entities that appeared in at least two different documents and were involved in more than two events.

The target entities are restricted to type PERSON (single persons or sets of people), ORGANISATION (corporations, agencies, and other groups of people defined by an established organisational structure), PRODUCT (anything that might satisfy a want or need, including facilities, food, products, services, etc.), and FINANCIAL (the entities belonging to the financial domain that are not included in one of the other entity types).

Some examples of target entities are *Steve Jobs* (PERSON), *Apple Inc.* (ORGANISATION), *Airbus A380* (PRODUCT), and *Nasdaq* (FINANCIAL).

The annotation procedure for the creation of gold standard timelines for the target entities required one person month. It consisted of four steps, as described below.

Entity annotation. All occurrences of the target entities in the four corpora were marked following (Tonelli et al., 2014). Cross-document co-reference was annotated according to the NewsReader cross-document annotation guidelines (Speranza and Minard, 2014). For this task, we used CROMER[3] (Girardi et al., 2014), a tool designed specifically for cross-document annotation.

[2] `http://en.wikinews.org`.

[3] `https://hlt.fbk.eu/technologies/cromer`

Event and time anchor annotation. Using CROMER, the corpora were annotated with events following the NewsReader cross-document annotation guidelines (Speranza and Minard, 2014). The annotation of events as defined in (Tonelli et al., 2014) was restricted by limiting the annotation to events that could be placed on a timeline. Thus, we did not annotate adjectival events, cognitive events, counter-factual events (which certainly did not happen), uncertain events (which might or might not have happened) and grammatical events[4]. For example, the events *gave*, *chosen* and *been (on medical leave)* in Figure 1 are excluded from the timeline as they are grammatical events.

Furthermore, timelines only contain events in which target entities explicitly participate in a *has_participant* relation as defined in (Tonelli et al., 2014), with the semantic role ARG0 (i.e. agent) or ARG1 (i.e. patient), as defined in the PropBank Guidelines (Bonial et al., 2010). In the example in Figure 1 we have an explicit *has_participant* relation between the entity *Steve Jobs* and the event *fighting* with semantic role ARG0, and one with semantic role ARG1 between *Steve Jobs* and *described*.

Based on TimeML (Pustejovsky et al., 2003), a time anchor corresponds to a TIMEX3 of type DATE; the time anchor attribute of an event takes as value the point in time when the event occurred (in the case of punctual events) or began (in the case of durative events). Its format follows the ISO-8601 standard: YYYY-MM-DD (i.e. Year, Month, and Day).

The finest granularity for time anchor values is DAY; other granularities admitted are MONTH and YEAR (references to months are specified as YYYY-MM and references to years are expressed as YYYY). The place-holder character, X, is used for unfilled positions in the value of a component. Thus, an event happened some day (not specified in the text) in July 2010 (for example, *resigned* in *The company's CEO met his employees one morning last July*) has time anchor 2010-07-XX (granu-

[4] Grammatical events are verbs or nouns that are semantically dependent on a governing content verb/noun. Typical examples of grammatical events are copula verbs, light verbs followed by a nominal event, aspectual verbs and nouns, verbs and nouns expressing causal and motivational relations, and verbs and nouns expressing occurrence.

larity DAY), while an event happened in the same month but with a granularity lower than day (for example in *Apple received criticism last month for the placement of the antenna on iPhone 4*), has time anchor 2010-07. Similarly, XXXX-XX-XX is used when the time anchor is completely unknown and the granularity is DAY, while XXXX-XX and XXXX are used when the time anchor is unknown and the granularity is MONTH and YEAR respectively (Minard et al., 2014a).

Automatic creation of timelines. We represent timelines in a simple tab format. On each line, we first have a cardinal number indicating the position of an event in the timeline, then the value of the anchor time attribute for the same event, and finally the event itself, which is represented as follows: document identifier, sentence number and textual extent of the event. For example, the event *18315-7-leave* in Figure 1 (occurring in sentence 7 of document 18315) occupies position *4* in the timeline and is anchored to *2011-01*.

In the case of event coreference, in the third column, there is a list of coreferring events separated by tabs instead of a single event (see the coreferring events *18315-7-fighting* and *18355-4-fighting* at position 1 in the example in Figure 1).

If two events have the same value for the anchor time attribute, they are placed in the same position (i.e. the same number in the first column), but on different lines.

The automatic created timelines are produced by a script that orders events in a timeline on the basis of the time anchors (all events with the same time anchor are simultaneous and all events with unknown time anchor are at position 0).

Manual revision of the timelines. The manual revision consists of ordering events with the same time anchor or with unknown time anchor taking into consideration textual information that goes beyond the defining of time anchor (Minard et al., 2014a).

For example both *founded* and *closed* in *The firm was founded in 2010 and closed before the end of the year* have anchor time 2010; nonetheless, based on textual information, it is possible to order them (the firm first was founded and then closed). When it is not possible to order events based either on the time anchor or on textual information, annotators leave them at the same position on the timeline. The same holds for events with anchor time XXXX-XX-XX; if annotators have no textual information that can help ordering them, they leave them at position 0; otherwise they place them on the timeline.

Inter-annotator agreement Three annotators have annotated a corpus starting from one target entity, i.e. they have annotated entity coreferences refering to the target entity and the events in which this entity participates. The corpus used is the trial corpus about *Apple Inc.* and the target entity *iPhone 4*. We compute the inter-annotator agreement using the Dice's coefficient (Dice, 1945). For the annotation of entity and event mentions, the agreement is respectively 0.81 and 0.66, and for entity coreferences of 0.84.

4 Task Dataset

The dataset used for this task is composed of articles from Wikinews, a collection of multilingual online news articles written collaboratively in a wiki-like manner. The reason for choosing Wikinews as a source is its creative commons license allowing us to freely release this dataset to the research community. For this task, we selected Wikinews articles around four topics:

- Apple Inc. (trial corpus);
- Airbus and Boeing (corpus 1);
- General Motors, Chrysler and Ford (corpus 2);
- Stock Market (corpus 3).

The trial data consists of one corpus of 30 documents and gold standard timelines for six target entities. The other three corpora, each consisting of 30 documents (about 30,000 tokens each) were used as the evaluation dataset.

As reported in Table 1, the total number of target entities in the evaluation dataset amounts to 38, but for the evaluation we used 37 timelines instead as one of the timelines contained no events.

The trial data contains one target entity of type ORGANISATION, one of type PERSON and 4 of type PRODUCT. The distribution of target entity types in the evaluation dataset is the following: 18 of type ORGANISATION, 10 of type FINANCIAL, 7 of type PERSON and 3 of type PRODUCT.

	Trial corpus	Evaluation dataset			
	Apple Inc.	Airbus	GM	Stock	Total
# documents	30	30	30	30	90
# sentences	464	446	430	459	1,335
# tokens	10,373	9,909	10,058	9,916	29,893
# events	187	343	308	264	915
# event chains	168	244	234	210	688
# target entities	6	13	12	13	38
# timelines	6	13	11	13	37
# events / timeline	31.2	26.4	25.7	20.3	24.1
# event chains / timeline	28	18.8	19.5	16.2	18.1
# docs / timeline	5.8	6.2	5.7	9.1	6.9

Table 1: Quantitative data about the dataset.

The three evaluation corpora are very similar in terms of size. It is interesting to notice, however, that the timelines created from the Stock Market corpus have peculiar features as they contain a lower average number of events with respect to those created from the other corpora. On the other hand, on average, Stock Market timelines contain events from a higher number of different documents, i.e. 9.1, versus 6.2 for Airbus and 5.7 for GM.

5 Evaluation Methodology

The evaluation methodology of this task is based on the evaluation metric used for TempEval-3 (UzZaman et al., 2013) to evaluate relations in terms of recall, precision and F_1-score. The metric captures the temporal awareness of an annotation (UzZaman and Allen, 2011).

> Temporal awareness is defined as the performance of an annotation as identifying and categorizing temporal relations, which implies the correct recognition and classification of the temporal entities involved in the relations.
>
> We calculate the Precision by checking the number of reduced system relations that can be verified from the reference annotation temporal closure graph, out of number of temporal relations in the reduced system relations. Similarly, we calculate the Recall by checking the number of reduced reference annotation relations that can be verified from the system output's temporal closure graph, out of number of temporal relations in the reduced reference annotation. (UzZaman et al., 2013)

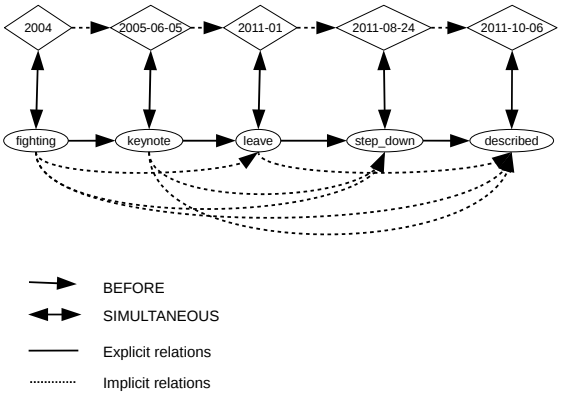


Figure 2: Explicit and implicit relations resulting from the timeline of Figure 1.

Before evaluating temporal awareness, each timeline needs to be transformed into a set of temporal relations. Figure 2 shows the explicit relations resulting from the timeline of Figure 1 as well as the implicit relations captured by the temporal graph. In order to convert each timeline, we defined the following transformation steps:

1. Each time anchor is represented as a TIMEX3.
2. Each event is related to one TIMEX3 with the SIMULTANEOUS relation type.

3. If one event happens before another one, a BEFORE relation type is created between both events.
4. If one event happens at the same time as another one, a SIMULTANEOUS relation type is created between both events.

Note that the evaluation of subtracks (ordering only), requires steps 3 and 4 alone.

For this first pilot on timelines, we decided to simplify the representation of durative events in the timelines by anchoring them in time considering their starting point. For this reason we represent relations between each event and its time anchor with the SIMULTANEOUS relation type (instead of other possibilities like BEGUN_BY or INCLUDES).

Events placed at the beginning of the timeline at position 0, i.e. events that were not ordered, are not considered in the evaluation. The official scores are based on the micro-average of the individual F_1-scores for each timeline, i.e. the scores are averaged over the events of the timelines of each corpus. The micro-average precision and recall values are also provided.

6 Participant Systems

29 teams signed up for the evaluation task, 8 teams downloaded the evaluation dataset and only 4 teams submitted results. A total of 13 unique runs were submitted: 3 for Track A (for which the participants worked on the raw texts), 2 for SubTrack A, 4 for Track B (for which the event mentions were provided) and 4 for SubTrack B.

The WHUNLP team processed the texts with Stanford CoreNLP. They applied a rule-based approach to extract target entities and their predicates, and perform temporal reasoning.

The SPINOZAVU[5] system is based on the pipeline developed in the NewsReader project and on the TIPSem tool. The tools are used for pre-processing, dependency parsing, semantic role labelling, event detection, temporal expression normalisation, coreference resolution and temporal relations extraction.

The GPLSIUA team also used a pipeline approach, employing the OpeNER language analysis toolchain, the Semantic Role Labeller from SENNA and the TIPSem tool for temporal processing. In addition, in order to detect event coreferences, they used the topic modelling algorithm of MALLET.

The HEIDELTOUL team used the HeidelTime tool for time expression recognition and normalisation and Stanford CoreNLP for coreference resolution. Afterwards, they used a cosine similarity matching function and a distance measure to select sentences relevant for a target entity and their events.

Three teams, SPINOZAVU, GPLSIUA and HEIDELTOUL, participated in the subtracks. They all submitted the same timelines both for the Tracks and the SubTracks, simply removing time anchors.

7 Evaluation Results

The official results are presented in Table 2. For each corpus we present the micro F_1-score and in the last three columns the micro precision, micro recall and micro F_1-score overall the three corpora. In the main track, Track A, WHUNLP_1 was the best run and achieved an F_1 of 7.28%. In Track B, GPLSIUA_1 obtained the best scores with an F_1 of 25.36%.

The subtracks were proposed in order to evaluate systems that do not perform time normalisation or event anchoring in time but focus on temporal relations between events. In the end, the events ordering of the runs submitted to the subtracks was the same as those submitted to the main tracks. In SubTrack A the best results are obtained with the run 1 of SPINOZAVU team, achieving an F_1-score of 1.69%. In SubTrack B, the best system is the same as in Track B, GPLSIUA_1, with an F_1-score of 23.15%.

We evaluate the selection of the relevant events involving a target entity using the classic evaluation metrics: recall, precision and F_1-score. All events are taken into account independently of their ordering in timelines; events placed at position 0 are also evaluated. The number of true positives and F_1-scores obtained on each corpus as well as the micro-average F_1-scores are presented in Table 3. In Table 3 we also provide the evaluation of time anchors assignment in terms of accurracy. For each timeline, the accurracy is computed by dividing the number of matching events/time anchors by the number of

[5]The members of the SPINOZAVU team involved in the NewsReader project were not involved in any annotation work or discussions around the organisation of the TimeLine task.

Track	Team run	Airbus F_1	GM F_1	Stock F_1	Total P	Total R	Total F_1
Track A	WHUNLP_1	8.31	6.01	6.86	14.10	4.90	**7.28**
	WHUNLP_1 [6]	*9.42*	*5.97*	*7.26*	*14.59*	*5.37*	***7.85***
	SPINOZAVU-RUN-1	4.07	5.31	0.42	7.95	1.96	3.15
	SPINOZAVU-RUN-2	2.67	0.62	0.00	8.16	0.56	1.05
SubTrackA	SPINOZAVU-RUN-1	1.20	1.70	2.08	6.70	0.97	**1.69**
	SPINOZAVU-RUN-2	0.00	0.92	0.00	13.04	0.14	0.27
TrackB	GPLSIUA_1	22.35	19.28	33.59	21.73	30.46	**25.36**
	GPLSIUA_2	20.47	16.17	29.90	20.08	26.00	22.66
	HEIDELTOUL_2	16.50	10.94	25.89	13.58	28.23	18.34
	HEIDELTOUL_1	19.62	7.25	20.37	20.11	14.76	17.03
SubTrackB	GPLSIUA_1	18.35	20.48	32.08	18.90	29.85	**23.15**
	GPLSIUA_2	15.93	14.44	27.48	16.19	23.52	19.18
	HEIDELTOUL_2	13.24	15.88	21.99	12.18	26.41	16.67
	HEIDELTOUL_1	12.23	14.78	16.11	19.58	11.42	14.42

Table 2: Official results of the TimeLine task of the four participating teams[7] presented per subcorpus and over the whole dataset. *(***Track A***: timelines with time anchors from raw text;* **SubTrack A***: timelines without time anchors from raw text;* **Track B***: timelines with time anchors from texts annotated with events;* **SubTrack B***: timelines without time anchors from texts annotated with events.)*

	Airbus			GM			Stock			Total		
	Events		TA	Events		TA	Events		TA	Events		TA
Team runs	TP	F_1	Acc	TP	F_1	Acc	TP	F_1	Acc	TP	F_1	Acc
WHUNLP	120	34.53	42.50	120	34.33	34.17	91	42.52	17.58	331	**36.33**	**32.63**
SPINOZAVU_1	46	17.59	23.91	61	22.93	36.07	57	30.24	0.00	164	22.91	20.12
SPINOZAVU_2	30	13.16	26.67	50	21.69	30.00	45	26.55	0.00	125	19.90	18.40
GPLSIUA_1	240	59.33	36.67	234	67.73	24.34	190	72.80	43.16	664	**65.68**	**34.17**
GPLSIUA_2	197	53.53	32.49	188	57.58	22.87	152	59.14	41.45	537	56.44	31.66
HEIDELTOUL_1	172	50.44	38.95	119	49.90	10.92	98	46.34	47.96	389	49.18	32.65
HEIDELTOUL_2	250	45.83	37.60	182	54.98	16.48	178	55.02	48.31	610	50.83	**34.43**

Table 3: Evaluation of the selection of events in which a target entity is involved and of time anchors assignment; *TP*: number of correctly identified events; F_1: micro-average F_1-score for the selection of events; *Acc*: accurracy in assignment of time anchors.

correctly identified events (TP in the table).

The results obtained in SubTracks, when evaluating only events ordering, are mainly lower than in Tracks, except on the "GM" corpus. For example the HEIDELTOUL_1 system achieved an F_1-score of 17.03% overall the 3 corpora in Track B and 14.42% in SubTrack B. But on "GM" corpus, the HEIDELTOUL_1 system obtained an F_1-score twice as high as in Track B, obtaining an F_1-score of 14.78% (vs. 7.25% in Track B). In evaluating the time anchors assignment (see Table 3), we observed that HEIDELTOUL and GPLSIUA systems performed better on the "Airbus" and "Stock" corpora than on "GM". This explains in part the better performance of their systems on the "GM" corpus when evaluating only events ordering (SubTrack B) than when evaluating both time anchors assignment and events ordering

[6] We found an error in the format of some event ids and re-processed the evaluation on a corrected version of the timelines.

[7] HEIDELTOUL_1 and HEIDELTOUL_2 are shorthand for HEIDELTOUL_NONTOLMATCHPRUNE and HEIDELTOUL_TOLMATCHPRUNE respectively.

(Track B). Furthermore, the task of time expression extraction and normalisation has been the topic of different shared tasks and the obtained results are high with an F_1-score of 90.30 for time expression detection and of 77.61 for normalisation (results obtained by HeidelTime (Strötgen et al., 2013) at TempEval-3). However, the performance of temporal relation extraction systems is quite low with an F_1-score of 36.26 obtained by ClearTK-2 (Bethard, 2013), the best system at TempEval-3 on Task C.

Observing the results by corpus in Table 2, we notice that, except for Track A, the best results are obtained on the "Stock Market" corpus. One of the reasons is that in the timelines related to this corpus all events were ordered (only one event was placed at position 0), while in "Airbus" and "GM" corpora less than 70% of the events were ordered.

In the "GM" corpus, one timeline was empty ("General Motors creditors"), i.e. the corpus does not contain any event that have this target entity as Arg0 or Arg1, therefore this timeline was removed from the evaluation. We observed that SPINOZAVU systems in Track A and GPLSIUA systems in Track B correctly returned an empty timeline, while WHUNLP created a timeline with 3 events in Track A and HEIDELTOUL_1 and HEIDELTOUL_2 produced a timeline containing respectively 32 and 78 events for this target entity in Track B.

Track B was proposed as a simplified task given that annotated texts with events were distributed to participants. Unfortunately no results from the same system run on both Tracks A and B were submitted, therefore, at the moment, we cannot evaluate the impact of pre-annotation of events.

8 Conclusion

The TimeLine task is the first task focusing on cross-document ordering of events. For this task, we have defined guidelines for cross-document annotation and for timeline creation, as well as annotated trial and evaluation datasets. The results submitted by four teams show much room for improvement. Obviously, timeline creation is a very challenging task which deserves more attention in future research.

Additionally, during the organisation of this task, many issues arose that provide interesting avenues of future research into timeline creation. Our three main issues concern durative versus punctual events, events without explicit time anchors and the relation between target entities and events. Below, we detail each of these questions.

Anchoring events in time. The ordering of an event in a timeline is based on the time when the event occurred. However, many events are durative events that have a starting point and/or an ending point. For the task, we decided to order durative events according to their starting points. We are investigating whether a new timeline format can be defined to represent the durative aspect of these events.

Events without explicit textual time anchor. We made the choice to include them in the timelines but not to evaluate them (events at position 0). The difficulty is to identify cases in which an event cannot be ordered in order to give instruction to annotators and systems. When ordering an event, should we take into consideration the information contained inside one document or inside one corpus, or could (should) we consider also background knowledge?

The relation between target entities and events. We chose to select events in which one target entity is explicitly involved in a participant relation. Amongst others, this rule excludes events involving a group of which a target entity is member. For example the event *received* in *The two companies have received $13.4 billion* (in which *the two companies* refers to General Motors and Chrysler) does not appear either in the "General Motors" timeline or in the "Chrysler" timeline. Considering also implicit *has_participant* relations would take the timeline task into the domain of complex entity relationships, but could possibly be interesting if considered in combination with taxonomy induction tasks.

With this TimeLine task, we aimed to take a step forward in the current state-of-the-art in cross-document coreference and temporal relation extraction. As organisers, we needed to come up with new ways of annotating and representing data. For the participating teams, the task meant that they needed to combine cutting-edge NLP technologies. This pilot task has shown us that the goal of automatic timeline extraction from raw text is challenging, but it has given us many more insights into what is possible, and what issues still need to be addressed.

Acknowledgments

This research was funded by the European Union's 7th Framework Programme via the NewsReader (ICT-316404) project.

References

Steven Bethard. 2013. ClearTK-TimeML: A minimalist approach to TempEval 2013. In *Proceedings of the Seventh International Workshop on Semantic Evaluation*, SemEval '13, pages 10–14, Atlanta, Georgia, USA, June.

Claire Bonial, Olga Babko-Malaya, Jinho D. Choi, Jena Hwang, and Martha Palmer. 2010. Propbank annotation guidelines, December. http://www.ldc.upenn.edu/Catalog/docs/LDC2011T03/propbank/english-propbank.pdf.

Lee Raymond Dice. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, July.

Christian Girardi, Manuela Speranza, Rachele Sprugnoli, and Sara Tonelli. 2014. CROMER: a Tool for Cross-Document Event and Entity Coreference. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may.

Heng Ji, Ralph Grishman, Zheng Chen, and Prashant Gupta. 2009. Cross-document event extraction and tracking: Task, evaluation, techniques and challenges. In *RANLP*, pages 166–172.

Anne-Lyse Minard, Alessandro Marchetti, Manuela Speranza, Bernardo Magnini, Marieke van Erp, Itziar Aldabe, Rubén Urizar, Eneko Agirre, and German Rigau. 2014a. TimeLine: Cross-Document Event Ordering. SemEval 2015 - Task 4. Annotation Guidelines. Technical Report NWR2014-11, Fondazione Bruno Kessler. http://www.newsreader-project.eu/files/2014/12/NWR-2014-111.pdf.

Anne-Lyse Minard, Manuela Speranza, Bernardo Magnini, Marieke van Erp, Itziar Aldabe, Rubén Urizar, Eneko Agirre, and German Rigau. 2014b. TimeLine: Cross-Document Event Ordering. SemEval 2015 - Task 4. Technical Report NWR2014-10, Fondazione Bruno Kessler. http://www.newsreader-project.eu/files/2013/01/SemEvaltaskdescription.pdf.

James Pustejovsky, José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003. TimeML: robust specification of event and temporal expressions in text. In *Fifth International Workshop on Computational Semantics (IWCS-5)*, pages 1–11.

Marta Recasens, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. SemEval-2010 Task 1: Coreference Resolution in Multiple Languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 1–8, Stroudsburg, PA, USA.

Manuela Speranza and Anne-Lyse Minard. 2014. NewsReader Cross-Document Annotation Guidelines. Technical Report NWR2014-9, Fondazione Bruno Kessler. http://www.newsreader-project.eu/files/2015/01/NWR-2014-9.pdf.

Jannik Strötgen, Julian Zell, and Michael Gertz. 2013. Heideltime: Tuning english and developing spanish resources for tempeval-3. In *Proceedings of the Seventh International Workshop on Semantic Evaluation*, SemEval '13, pages 15–19, Atlanta, Georgia, USA.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813, September.

Sara Tonelli, Rachele Sprugnoli, Manuela Speranza, and Anne-Lyse Minard. 2014. NewsReader Guidelines for Annotation at Document Level. Technical Report NWR2014-2-2, Fondazione Bruno Kessler. http://www.newsreader-project.eu/files/2014/12/NWR-2014-2-2.pdf.

Naushad UzZaman and James Allen. 2011. Temporal evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 351–356, Portland, Oregon, USA.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations. In *Proceedings of the Seventh International Workshop on Semantic Evaluation*, SemEval '13, pages 1–9, Atlanta, Georgia, USA.

Marc Verhagen, Robert J. Gaizauskas, Frank Schilder, Mark Hepple, Jessica Moszkowicz, and James Pustejovsky. 2009. The TempEval challenge: identifying temporal relations in text. *Language Resources and Evaluation*, 43(2):161–179.

Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 Task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 57–62, Stroudsburg, PA, USA.

3 Computing News Story Workshop

In the scope of the NewsReader project we took part in the proposition and organization of the First Worskop on Computing News Storylines (NewsStory 2015) which was held at ACL 2015 on July 31, 2015.

The NewsStory workshop aims at bringing together researchers and scientists working on narrative extraction from news and its representation. Narratives are at the heart of information sharing. Ever since people began to share their experiences, they have connected them to form narratives. Modern day news reports still reflect this narrative structure, but they have proven difficult for automatic tools to summarize, structure, or connect to other reports. The workshop aims at assessing state-of-the-art event extraction and linking, as well as detecting and ranking narratives according to salience.

The participants are encouraged to re-use the SemEval 2015 TimeLine Task datasets to provide their own annotations, interpretations, and system results. The data will be collected and summarized before the workshop to facilitate an insightful comparison. The results of the combined manual and automatic annotation of the common dataset will be used to dive the discussion around three themes:

- Definitions: what is a storyline? how can it be formally and computationally formulated?
- Resources: what are the core markables of a storyline? how should annotation of storylines be performed? can existing annotation schemes be re-used and adapted for storyline annotation? how should we annotate cross-document information concerning events and character perspectives? is it feasible to develop a StoryBank for evaluation?
- Evaluation: how do we determine if an extracted storyline is "good enough"? can standard measures, such as Precision, Recall and F-measure, be applied to evaluate storyline extraction or do we need different measures? should evaluation take place at a global level or must it be conducted separately on the different components of a storyline system?

We received 12 submissions: 5 long papers and 7 short papers. We selected 9 papers, including 4 with conditional acceptance. The workshop program can be found at `https://sites.google.com/site/computingnewsstorylines2015/program`. The proceedings of the workshop are published in the ACL Anthology at `https://aclweb.org/anthology/W/W15/W15-45.pdf`.

The important dates of the workshop are the following:

- First Call for Papers: January 11, 2015
- Second Call for Papers: February 11, 2015
- Deadline for Submissions: May 21, 2015

- Author Notification: June 4, 2015
- Camera-ready version: June 21, 2015
- Workshop: July 31, 2015

The Organizing Committee is made up of the following: Tommaso Caselli, Marieke van Erp, Anne-Lyse Minard, Mark Finlayson, Ben Miller, Jordi Atserias, Alexandra Balahur and Piek Vossen.

A website containing all the information related to the workshop was set up: `https://sites.google.com/site/computingnewsstorylines2015/`.

A second edition of the workshop will be organized in conjunction with EMNLP 2016 in Austin, TX, USA in November 2016.

4 Evalita Task: EVENTI

Through FBK, the NewsReader project endorsed the "EValuation of Events aNd Temporal Information" (EVENTI [2]) task at Evalita 2014, an initiative devoted to the Natural Language Processing and Speech tools for Italian. EVENTI focused on the temporal processing of Italian texts following the TempEval SemEval task. It consisted of a Main task on contemporary news and a Pilot task on historical texts and included four subtasks:
A. Temporal Expression recognition and normalization;
B. Event identification and classification;
C. Identification of temporal relations from raw text;
D. Classification of temporal relation holding between a given pair of elements (either event/event or event/timex pairs).

The task started on March 15, 2014, when the Task guidelines and development data were made available, and the evaluation took place between September 8, 2014 (when test data were made available) and September 22, 2014 (when assessment was returned to participants). Although eight teams registered for the task, only three actually submitted the output of their systems, for a total of 17 unique runs. The final results and official ranking were made public at the Evalita 2014 final workshop, which was held on December 11, 2014, in Pisa, within the XIII AI*IA Symposium on Artificial Intelligence.

The materials provided to the participants can be accessed through the workshop's website: `https://sites.google.com/site/eventievalita2014/`.

The Task was organized by Tommaso Caselli, Rachele Sprugnoli, Manuela Speranza and Monica Monachini.

In the remainder of the section we have included the paper published in the proceedings of the Evalita workshop.

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

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 Information*) 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.

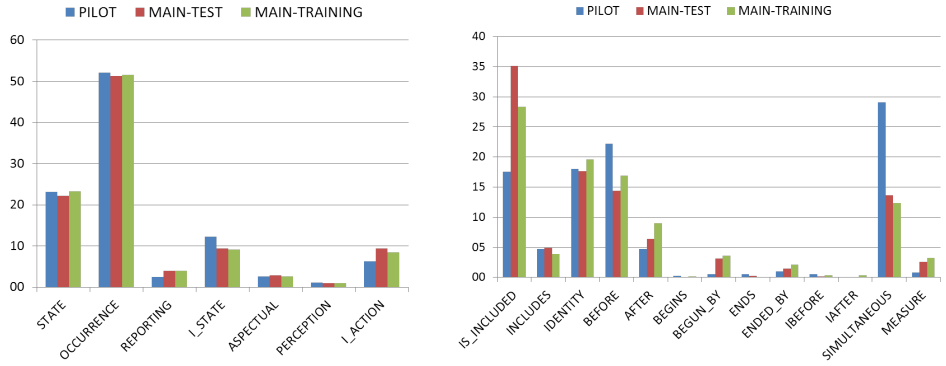
- Subtask D: determine the value of the temporal relation given two gold temporal elements (i.e. the source and the target of the relation) as defined in Task C (main event - main event; main event - subordinate event; event - timex).

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.

[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/

	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

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 Ital-

ian 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.

5 The first CLIN Dutch Shared Task

VUA organized the first CLIN Dutch Shared Task, which was part of CLIN26. We used the Dutch section of the NewsReader MEANTIME corpus. Participants were invited to join in one or more of several subtasks for which the annotations are present in the MEANTIME corpus. The following subtasks were included in the task (all tasks were given a gold set of data annotated according to the NewsReader annotation guidelines for Dutch Schoen *et al.* (2014)):

- Named Entity Recognition and Classification. This task involves the MEANTIME entity annotations, which are both evaluated on the inner spans and outer spans. The outer spans correspond to the ones occurring in the CoNNL data set (Tjong Kim Sang and De Meulder (2003)), inner spans are smaller spans of named entities that are embedded in the span of a larger named entity.

- Nominal coreference. This task consists of identifying which expressions in the text refer to the same entity.

- Event recognition and factuality. In this task, all events mentioned in the text must be identified and for each event, it should be determined whether it is:

  - certain, probable, possible (or underspecified in certainty)
  - affirmed or denied (or underspecified), indicated as positive or negative polarity

The idea behind this shared task is that, by proposing a wide variety of tasks, several groups working on Dutch Natural Language Processing will be encouraged to join in for at least one or two tasks they have worked on before and may possibly try and tackle a new task.

The planning for this overall Dutch shared task was as follows:

- Development data release: September 9, 2015

- Evaluation data release: November 15, 2015

- System output due: November 30, 2015

- Results and presentations: December 18, 2015 (during CLIN26)

In order to create the development and evaluation data for each task, the annotated files of the MEANTIME corpus were converted from XML stand-off format to CoNLL format. We split the data in two sets and provided participants with 30 documents as development data and 90 documents as evaluation data.

The following teams participated in the task (we present the abstracts provided by the participants):

Groref. *Rule-Based Coreference Resolution for Dutch* (Rob van der Goot, Hessel Haagsma and Dieke Oele): We have adapted Stanford's multi-pass sieve coreference resolution system for Dutch. Our experiments prove that this rule-based system works robustly on Dutch, for different domains. Because no training data is needed, it is a well-suited approach for low-resource languages.

Languagemachines. *Running Frog on the CLIN26 NER task* (Iris Hendrickx, Ko van der Sloot, Maarten van Gompel and Antal van den Bosch): Frog is an integration of memory-based natural language processing (NLP) modules developed for Dutch. All NLP modules are based on Timbl, the Tilburg memory-based learning software package. Most modules were created in the 1990s at the ILK Research Group (Tilburg University, the Netherlands) and the CLiPS Research Centre (University of Antwerp, Belgium). Over the years they have been integrated into a single text processing tool, which is currently maintained and developed by the Language Machines Research Group at Radboud University Nijmegen. For the CLIN26 NER task we applied the Frog named-entity recognizer module that is trained on the SoNaR1 named entity labels. We will briefly discuss the architecture of Frog, its NER module, and the results. The main source of misclassifications is the difference between the label sets of SoNar1 and the CLIN task.

Irismonster. *Rule based classification of events and factuality* (Iris Monster, Iris Hendrickx-Dekkers): The aim of this project was to solve the event detection and factuality classification task of CLIN26. The data was parsed by an annotation tool called Frog. Using these annotations (Part of Speech-tags, lemma, et cet.) a rule-based classifier was built that discovers events. In order to annotate the resulting events with corresponding polarity and factuality tags two other rules-based classifiers were implemented. The performance of the classifiers varied on the test corpora. This could be due to the small training set.

RuGGED. *Event detection and event factuality classification for shared task* (Oliver Louwaars and Chris Pool): We present RuGGED, our approach to detecting and events and determine their factuality. Considering the little amount and high skewness of the data provided to develop our system, we opted for a rule-based rather than learning approach. Rule development was heavily based on the annotation guidelines and on data observation.

The results can be found in Table 1.

The development data, evaluation data, and all scorers have been made available at `https://github.com/cltl/clin26-eval`. All participants mentioned above presented their results at CLIN26. The Shared Task session attracted a full room (around 50 researchers), several of which expressed the wish for new shared task to become a new tradition. The first CLIN shared task is indeed intended to become the first of a series of CLIN shared tasks.

6 Italian Shared Task proposal

The data annotated within the project for Italian will be used as test data for a shared task that will be proposed at Evalita 2016.

System	Task	Precision	Recall	F-score
GroRef	Entity Detection	55.32	62.96	58.01
GroRef	Coreference (BLANC)	30.30	26.72	25.04
Languagemachines	Named Entity Recognition	56.20	52.10	47.80
RuGGed	Event Detection	63.80	77.50	68.97
Irismonster	Event Detection	71.17	48.63	56.50
RuGGed	Event Certainty	95.10	97.30	96.10
Irismonster	Event Certainty	92.90	74.20	81.90
RuGGed	Event Polarity	87.20	89.40	88.20
Irismonster	Event Polarity	88.20	70.50	77.80

Table 1: Results from the CLIN26 shared task

FBK is planning of exploiting the Italian section of the MEANTIME corpus as test data for a shared evaluation task. The task, to be proposed after the official closing of the project (most likely, at Evalita 2016), will consist of two subtasks: the first about time anchoring and the second one about event factuality.

- Time anchoring: Given a set of gold events, participant systems are required to detect those for which it is possible to identify a time anchor. Our definition of time anchor includes two different types of elements, i.e. the temporal expressions occurring in the text, as well as the Document Creation Date that is part of metadata associated to each document.
- Event factuality: For each event mention provided, participant systems are required to assign values for three attributes: certainty, polarity, and time.

This task is a continuation of the EVENTI-Evalita 2014 task on temporal processing for Italian.

The task has been defined and presented in a paper at CLiC-it 2015 in the track "Towards EVALITA 2016: challenges, methodologies and tasks":
Anne-Lyse Minard, Manuela Speranza, Rachele Sprugnoli and Tommaso Caselli. "FacTA: Evaluation of Event Factuality and Temporal Anchoring." In *Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it 2015)*, 2015.

7 Conclusions

In this deliverable, we give an overview of three evaluation tasks that we have organized in open competitions. The SemEval task focused on cross-document timeline creation in

English. It was the first evaluation task that addressed the challenging task of timeline creation. The results obtained by the participants show much room for improvement. Moreover this task sets up a first framework for evaluation timeline creation systems.

The NewsReader project also supported the organization of a task for Italian on temporal processing: EVENTI at Evalita 2014. This task was in the line of temporal processing tasks already organized for English and Spanish in 2007, 2010 and 2013 at SemEval.

More recently the first Dutch Shared Task has been organized at CLIN26. It included three tasks: Named Entity Recognition and Classification, Nominal Coreference, and Event Recognition and Factuality.

Furthermore, we have presented a workshop on Computing News Stories that was held at ACL in July 2015. The workshop gave us the opportunity to lead a discussion about the definition of a storyline, the annotation of storylines and the evaluation of storyline extraction. A second edition of the workshop will take place in November 2016.

By way of three evaluation tasks and a workshop, the NewsReader project has carried out the creation of evaluation frameworks for time processing of news with the further goal of evaluating complex storyline extraction. The project provides also evaluation frameworks for other languages than English: for Italian, a shared task was held at Evalita 2014 and a new one will be proposed for Evalita 2016; for Dutch a shared task was organized at CLIN26 in 2015.

References

Anneleen Schoen, Chantal van Son, Marieke van Erp, and Hennie van Vliet. Newsreader document-level annotation guidelines-dutch techreport 2014-8. Technical report, VU University, 2014.

Erik F Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics, 2003.

Marieke van Erp, Piek Vossen, Rodrigo Agerri, Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Egoitz Laparra, Itziar Aldabe, and German Rigay. Anntotated Data, version 2. Technical Report D3-3-2, VU Amsterdam, 2015. `http://www.newsreader-project.eu/files/2012/12/NWR-D3-3-2.pdf`.