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:

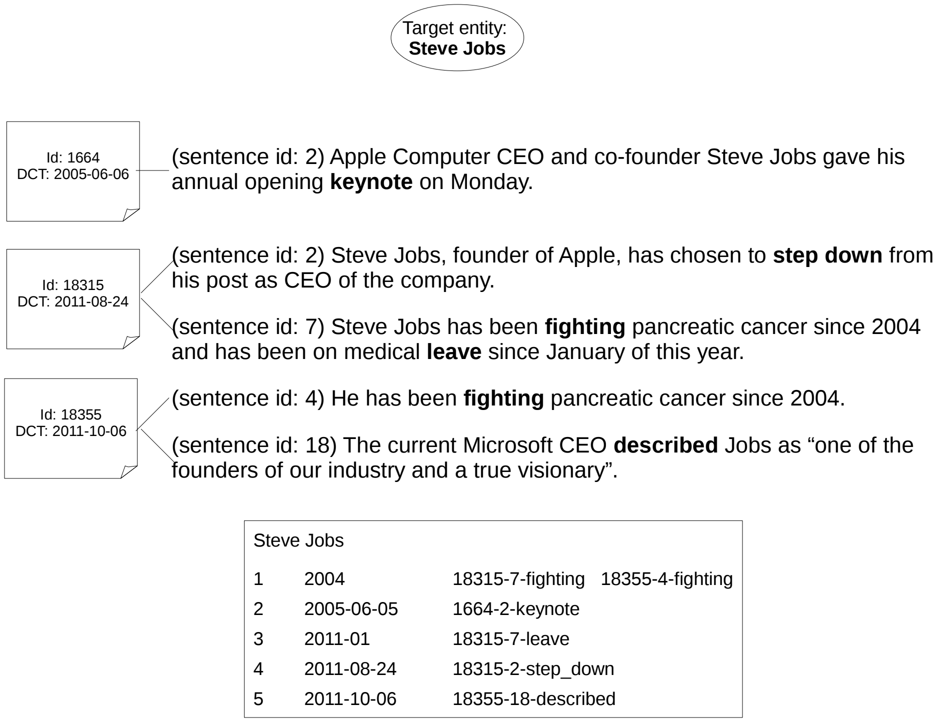
- TempEval-1 (2007): Temporal Relation Identification (Verhagen et al., 2009)
- TempEval-2 (2010): Evaluating Events, Time Expressions, and Temporal Relations (Verhagen et al., 2010)
- TempEval-3 (2013): Temporal Annotation (UzZaman et al., 2013)

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

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 778–786,
Denver, Colorado, June 4-5, 2015. ©2015 Association for Computational Linguistics



(sentence id: 2) Apple Computer CEO and co-founder Steve Jobs gave his annual opening **keynote** on Monday.

(sentence id: 2) Steve Jobs, founder of Apple, has chosen to **step down** from his post as CEO of the company.

(sentence id: 7) Steve Jobs has been **fighting** pancreatic cancer since 2004 and has been on medical **leave** since January of this year.

(sentence id: 4) He has been **fighting** pancreatic cancer since 2004.

(sentence id: 18) The current Microsoft CEO **described** Jobs as "one of the founders of our industry and a true visionary".

Steve Jobs			
1	2004	18315-7-fighting	18355-4-fighting
2	2005-06-05	1664-2-keynote	
3	2011-01	18315-7-leave	
4	2011-08-24	18315-2-step_down	
5	2011-10-06	18355-18-described	

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.

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-

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

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

[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)

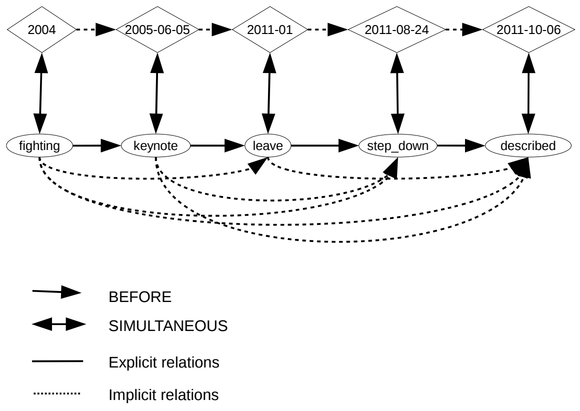


Figure 2: Explicit and implicit relations resulting from the timeline of Figure 1.

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.

		Airbus	GM	Stock	Total		
Track	Team run	F_1	F_1	F_1	P	R	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 reprocessed 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.