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

Technical Report NWR-2014-10
Version FINAL

Anne-Lyse Minard, Manuela Speranza, Bernardo Magnini
Fondazione Bruno Kessler

Marieke van Erp
VU University Amsterdam

Itziar Aldabe, Rubén Urizar, Eneko Agirre, German Rigau
The University of the Basque Country

BUILDING STRUCTURED EVENT INDEXES OF LARGE VOLUMES OF FINANCIAL AND ECONOMIC DATA FOR DECISION MAKING

ICT 316404

Table of Contents

1. Introduction

In any domain, professionals need to have access to knowledge in order to take well-informed decisions. An insightful way of presenting information in an easily updatable and complete manner is to present it on a timeline that is continuously updated with new information. The aim of the task is to build timelines from written news in English. More specifically, the goal is to order on a timeline all the events in which a target entity is involved. We focus mainly on cross-document event coreference resolution and cross-document temporal relation extraction.

Temporal relation extraction has been the topic of the three past TempEval tasks as part of SemEval:

- TempEval-1 (2007): Temporal Relation Identification
- TempEval-2 (2010): Evaluating Events, Time Expressions, and Temporal Relations
- TempEval-3 (2013): Temporal Annotation

In addition, temporal relation extraction has been the focus of the 6th i2b2 NLP Challenge for clinical records (Sun et al., 2013).

The cross-document aspect, on the other hand, has not been often explored. Ji et al. (2009) worked on a similar task using the ACE 2005 training corpora. The task was to 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.

Partially motivated by the work in the NewsReader project,[1] TimeLine goes beyond the above mentioned tasks by addressing coreference resolution for events and temporal relation identification at a cross document level.

2. Task Description

Given a set of documents and a target entity, the task is to build an event TimeLine related to that entity, i.e. to detect, anchor in time and order the events involving the target entity.

Input data. As input data, we provide a set of documents and a set of target entities (people, organization, product or financial entity); only entities of interest will be selected as target entities, i.e. entities involved in many events across different documents and for which it is relevant to build a timeline.

Tracks. Two different tracks are proposed on the basis of the data used as input. For Track A only raw texts are provided to the participants, while for Track B gold event mentions are also given.
For both tracks the expected output is one TimeLine for each target entity. Each TimeLine consists of an ordered list of events in which each event is associated to a time anchor.

For both tracks a sub-track in which the events are not associated to a time anchor is proposed.

Track A (main track):
- input data: raw texts
- output: full TimeLines (ordering of events and assignment of time anchors)

Subtrack A:
- input data: raw texts
- output: TimeLines consist of just ordered events (no assignment of time anchors)

Track B:
- input data: texts with manual annotation of event mentions
- output: full TimeLines (ordering of events and assignment of time anchors)

Subtrack B:
- input data: texts with manual annotation of event mentions
- output: TimeLines consist of just ordered events (no assignment of time anchors)

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

Participants can choose to participate in any track and subtrack.
Participants can submit up to two runs for each track/subtrack.

TimeLine. A TimeLine is represented in a simple tab format:
ordering time_anchor event(s)

The first column (*ordering*) contains a cardinal number which indicates the position of the event in the TimeLine (two events can be associated to the same number if they are simultaneous). The second column (*time_anchor*) contains a time anchor. The third column (*event*) consists of one event or a list of corefered events separated by a tab. Each event is represented by the id of the file (<DOCID>), the id of the sentence and the extent of the event mention in the following format: docid-sentid-event (11778-2-launch)
In the case of multi-words event, tokens are separated by an underscore:
16844-12-showed_off

TimeLine example:
iTunes
1 2003 11778-3-launch 11778-4-launch
2 2007 11778-4-pass
3 2008-01 11778-7-hold
4 2008-02 11778-2-pass 11778-5-pass
4 2008-02 11778-3-accounts_for

Target Entities. Each TimeLine is associated to one target entity. The entity can be of type *organization*, *person*, *product* or *financial entities*.
The TimeLine contains events in which the target entity **explicitly** participates in a has_participant relation, according to the NewsReader Guidelines (section 10.2), with the semantic role ARG0 (i.e. agent) or ARG1 (i.e. patient), according to PropBank Guidelines (Bonia et al., 2010). In the sentence (1) *Iphone 4* is ARG0 of the verb *use*, and in sentence (2) it is ARG1 of the verb *unveil*.

(1) ***The iPhone 4*** *will use iOS.*
(2) *Yesterday, Steve Jobs unveiled* ***iPhone 4****.*

Entity coreference must be resolved. A TimeLine should contain events involving besides the target entity its coreferences (including pronominal coreferences). For example in a TimeLine about “Cook”, both events involving “Cook” in the first sentence and “He” in the second should be part of the TimeLines.
(3) Before his post at Apple, **Cook** held positions at IBM and Compaq. **He** is known for staying out of the spotlight.

The **member_of** relations are not considered as coreferences.

In sentence (4) “the parties” refers to the two companies “Apple Inc.” and “Apple Corps”, but “the parties” doesn’t corefer with neither “Apple Inc.” or “Apple Corps”.
(4) *On September 21, 2004* ***the parties*** *agreed to have the case heard by the UK court.*

Events. Not all events can be part of a TimeLine, amongst others counter-factual events will not appear in a TimeLine. The Manual Annotation Guidelines provides details about candidate events for TimeLines.

Event coreference must be resolved. For two coreferring events there is only one position (i.e. one line) in the TimeLine.
The sentence (5) and (6) contain two event mentions which corefer: “introduced” and ‘introducing”.
They will appear at the same position in the TimeLine:
1 2010-06-07 16844-5-introducing 16900-11-introduced

(5) *The newest iPhone, [iPhone 4] was* ***introduced*** *by [Apple CEO Steve Jobs] at the company's 2010 Worldwide Developer's Conference less than two weeks ago*.
(6) *While* ***introducing*** *[iPhone 4], at the annual conference, [Jobs]* [...]

Time Anchors. In a TimeLine each event is associated to a time anchor and the annotation of time anchors is based on TIMEML.
A time anchor is always a DATE (as defined in TIMEML) and its format follows the ISO-8601 standard: YYYY-MM-DD (that is Year, Month, and Day), the maximum granularity admitted being DAY.
As for anchors with a lower granularity, we admit only months and years: references to months are specified as: YYYY-MM, whereas references to years are expressed as: YYYY.
The place-holder character, X, is used for each unfilled position in the value of a component.

Examples:

- February 6, 2007 → 2007-02-06
- April 2010 → 2010-04
- in 2009 → 2009
- May 23 → XXXX-05-23

A time anchor takes as value the point in time when the event occurred (in case of punctual events) or began (in case of durative events).

Ordering. Event ordering is based on temporal relations between events; more specifically on the before/after and includes/simultaneous relations as defined by ISO-TimeML.

3. Examples

In this section, we give two examples of the task. In the examples we give excerpts of the documents associated to the document creation time (DCT), information available in each document. The events in which the target entity participates with the semantic role ARG0 (i.e. agent) or ARG1 (i.e. patient) are in bold. The produced TimeLine is given, with the anchor time and the order of the events.

1. Given the entity *Steve Jobs* as an input and a set of documents, a TimeLine is built.

- Entity: Steve Jobs
- Relevant sentences:
    1. (file id: 1664; DCT: June 6, 2005; sentence id: 2) *Apple Computer CEO and co-founder Steve Jobs gave his annual opening **keynote** to the World Wide Developers Conference (WWDC) at Moscone Center in San Francisco, California on Monday*.
    2. (file id: 18315; DCT: August 24, 2011; sentence id: 2) *Steve Jobs, founder of Apple, has chosen to **step down** from his post as CEO of the company.*
    3. (file id: 18315; DCT: August 24, 2011; sentence id: 7) *Steve Jobs has been **fighting** pancreatic cancer since 2004 and has been on medical **leave** since January of this year.*
    4. (file id: 18355; DCT: October 6, 2011; sentence id: 4) *He has been **fighting** pancreatic cancer since 2004.*

- TimeLine:

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	

2. For the second example, the entity of interest is *Beatles' Apple Corps*.

- Entity: *Beatles' Apple Corps*
- Relevant sentences:
    1. (file id: 4954; DCT: May 8, 2006; sentence id: 2) *The Beatles' label Apple Corps **lost** its court case against Apple Computer today in the High Court.*

2. (file id: 4954; DCT: May 8, 2006; sentence id: 6) *During the case Apple Corps **showed** the court just how many times the Apple Computer logo appeared during a typical download.*
3. (file id: 4596; DCT: March 28, 2006; sentence id: 3) *Apple Corps **claims** that Apple Computer's iTunes Music Store violates an agreement reached between the two companies in 1991.*

- TimeLine:

Beatles Apple Corps

1	2006-03-28	4596-3-claims
2	XXXX-XX-XX	4954-6-showed
3	2006-05-08	4954-2-lost

4. Data

Trial data

The trial data consists of a set of 30 documents collected from Wikinews (http://en.wikinews.org) about *Apple Inc*. A set of target entities (input) and the corresponding ordered list of events (the output timeline) is provided with the set of documents.

The trial data have been annotated with the extents of event mentions.

No training corpus will be provided in addition to the development corpus.

The full annotation following the NewsReader Guidelines of the 5 first sentences of 20 documents of the trial data is available on the NewsReader web site:
http://www.newsreader-project.eu/results/data/

Evaluation data

The evaluation data will consist of 3 sets of documents annotated with event mentions and a set of target entities. Each set will contain around 30 documents from Wikinews, totalling around 30,000 tokens.

5. Format

Documents

The documents will be available in two formats: CAT (Content Annotation Tool) labelled format[2] (Bartalesi Lenzi et al.,2012) and a format which mimics the TimeML format.[3]

[2]

CAT labelled format is an XML based standoff format where different annotation layers are stored in separate document sections and are related to each other and to source data through pointers. Trial data are annotated with event mentions and the document creation time, so each document contains 2 different sections: one with the tokens and one with the markables.

```
<Document doc_id="16900" doc_name="16900-Apple_swamped_by_iPhone">
  <token number="0" sentence="0" t_id="1">Apple</token>
  <token number="1" sentence="0" t_id="2">swamped</token>
  <token number="2" sentence="0" t_id="3">by</token>
  <token number="3" sentence="0" t_id="4">iPhone</token>
  <token number="4" sentence="0" t_id="5">4</token>
  <token number="5" sentence="0" t_id="6">pre-orders</token>
  <token number="6" sentence="1" t_id="7">June</token>
  <token number="7" sentence="1" t_id="8">16</token>
  <token number="8" sentence="1" t_id="9">,</token>
  <token number="9" sentence="1" t_id="10">2010</token>
  <token number="10" sentence="2" t_id="11">Pre-orders</token>
  <token number="11" sentence="2" t_id="12">of</token>
  <token number="12" sentence="2" t_id="13">the</token>
  <token number="13" sentence="2" t_id="14">recently</token>
  <token number="14" sentence="2" t_id="15">announced</token>
  <token number="15" sentence="2" t_id="16">iPhone</token>
  <token number="16" sentence="2" t_id="17">4</token>
```

Excerpt of the token layer of a CAT labelled format document.

```
 −<Markables>
  −<EVENT_MENTION m_id="80">
      <token_anchor t_id="100"/>
    </EVENT_MENTION>
  −<EVENT_MENTION m_id="81">
      <token_anchor t_id="113"/>
    </EVENT_MENTION>
  −<EVENT_MENTION m_id="82">
      <token_anchor t_id="112"/>
    </EVENT_MENTION>
  −<TIMEX3 functionInDocument="CREATION_TIME" m_id="43" type="DATE" value="2010-06-16">
      <token_anchor t_id="7"/>
      <token_anchor t_id="8"/>
      <token_anchor t_id="9"/>
      <token_anchor t_id="10"/>
    </TIMEX3>
  </Markables>
</Document>
```

Excerpt of the markable layer of a CAT labelled format document.

All the files can be uploaded and opened in CAT.
To obtain username and password to access the tool, please go to the following URL:

[3] http://timeml.org/site/publications/specs.html

http://dh.fbk.eu/resources/catcontentannotationtool

In the **alike TimeML format** events are annotated using only the EVENT element (and not the MAKEINSTANCE as in TimeML). Elements has been added to mark out the paragraphs (p), the sentences (s) and associate each sentence to an unique id. The text is tokenized.

```
−<TimeML>
  <DOCID>1514</DOCID>
 −<DCT>
    <TIMEX3 functionInDocument="CREATION_TIME" tid="t0" type="DATE" value="2005-05-07">2005-05-07</TIMEX3>
  </DCT>
  <EXTRAINFO>Reactions to Apple 's OS X Tiger</EXTRAINFO>
 −<TEXT>
    <s id="1">May 7 , 2005</s>
   −<s id="2">
      In chatrooms and on bulletin boards , Macintosh users and the Macintosh - curious are
      <EVENT eid="50">buzzing</EVENT>
      about Tiger , the newest version of Apple Computer 's Mac OS X , version 10.4 .
    </s>
```

TimeLine

One file by TimeLine must be created. The first line contains the target entity.
The name of the files must be the mention of the target entity in lower case, and the extension ".txt". In the case of multi-words entity, tokens will be separated by an underscore.
E.g.: steve_jobs.txt

Set of target entities

For each set of documents, one file is provided containing the list of target entities, one by line.

6. Evaluation Methodology

Participants will submit the TimeLines produced by their system for all target entities. Systems will be ranked based on the temporal awareness (UzZaman and Allen, 2011).

7. References

Bartalesi Lenzi Valentina, Moretti Giovanni, and Sprugnoli Rachele (2012). CAT: the CELCT Annotation Tool. In Proceedings of LREC 2012 (pp. 333338).

Heng Ji, Ralph Grishman, Zheng Chen, and Prashant Gupta. Cross-document event extraction and tracking: Task, evaluation, techniques and challenges. In *Proceedings of the International Conference RANLP-2009*, pages 166–172, Borovets, Bulgaria, September 2009. Association for Computational Linguistics.

Naushad UzZaman, Hector Llorens, James F. Allen, Leon Derczynski, Marc Verhagen, and James Pustejovsky. Tempeval-3: Evaluating events, time expressions, and temporal relations. *CoRR*, abs/1206.5333, 2012.

Naushad UzZaman and James Allen. (2011). Temporal Evaluation. In Proceedings of The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *JAMIA*, 20(5):806–813, 2013.

ISO TimeML Working Group. ISO TC37 draft international standard DIS 24617-1, August 14 2008. http://semantic-annotation.uvt.nl/ISO-TimeML-08-13-2008-vankiyong.pdf.

Claire Bonial, Olga Babko-Malaya, Jinho D. Choi, Jena Hwang, and Martha Palmer. Propbank annotation guidelines, version 3.0. Technical report, Center for Computational Language and Education Research, Institute of Cognitive Science, University of Colorado at Boulder, Pisa, Italy, 2010. http://clear.colorado.edu/compsem/documents/propbank_guidelines.pdf.