

NewsReader Guidelines for cross-document Annotation

NWR-2014-9

Version FINAL

Manuela Speranza, Anne-Lyse Minard
manspera, minard@fbk.eu

(1) Fondazione Bruno Kessler
Via Sommarive 18, 38123, Povo (Trento), Italy



BUILDING STRUCTURED EVENT INDEXES OF LARGE VOLUMES OF FINANCIAL
AND ECONOMIC DATA FOR DECISION MAKING
ICT 316404

Contents

1	Overview of annotation guidelines	4
2	Entity	4
2.1	Entity mentions	4
2.2	Entity instances	4
3	Event	5
3.1	Event mentions	5
3.2	Event instances	8
4	Annotation task	10
4.1	Phase 1: Entity annotation	10
4.2	Phase II: Event annotation	10

1 Overview of annotation guidelines

This document presents the annotation guidelines defined in the NewsReader project for the cross-document annotation task.

Annotation consists of two main tasks:

- cross-document entity coreference annotation of all entities annotated in the first 6 sentences of each file;
- cross-document entity and event coreference annotation starting from a set of seed entities. All the mentions that corefer to the seed entities should be annotated as well as the events in which the seed entities are participants.

For the identification of entity mentions, for the definition of their extent, and for the annotation of coreference, annotators should refer to the guidelines provided for the intra-document annotation task[Tonelli *et al.*, NWR2014-2-2]¹ if not specifically stated below. For the identification of event mentions, specific guidelines have been defined. They are described in Section 3.1. In this document we refer to the intra-document annotation guidelines as *NWR-2014-2-2*.

Sections 2 and 3 describe the annotation of entities and events, both the mentions and the instances.

In the section 4 we detail the annotation task, i.e. the different steps of the cross-document annotation.

2 Entity

2.1 Entity mentions

As far as the extent of entity mentions is concerned, annotators should apply the same guidelines provided for the intra-document annotation (for entity mention extent, see section 3.1 of NWR-2014-2-2). The only exception is that CONJ-mentions (i.e. entity mentions connected by a coordinating conjunction) will not be annotated. Note: Mentions do not have any attribute.

2.2 Entity instances

Each entity instance has the following attributes:

¹<http://www.newsreader-project.eu/files/2014/12/NWR-2014-2-2.pdf>

- class (compulsory): to be filled according to the guidelines provided for the intra-document annotation (see section 2.1 of NWR-2014-2-2 on Entity types);
- name (compulsory): to be filled according to the guidelines provided for the intra-document annotation (see section 2.4 of NWR-2014-2-2 on Tag Descriptor);
- short description (compulsory): a short description of the entity instance, whose aim is to distinguish it from other entity instances with similar names;
- external reference (compulsory): to be filled according to the guidelines provided for the intra-document annotation (see section 2.3 of NWR-2014-2-2);
- comment (optional).

3 Event

3.1 Event mentions

The annotation of event mentions is based on the intra-document annotation guidelines (for event mention extent, see section 5.1 of NWR-2014-2-2). Note: Mentions do not have any attribute.

“Event is used as a cover term to identify “something that can be said to obtain or hold true, to happen or to occur” [ISO TimeML Working Group, 2008]. This notion can also be referred to as eventuality [Bach, 1986] including all types of actions (punctuals or duratives) and states as well.”

Some events annotated following the NewsReader guidelines could not go on a timeline, for example because they didn’t happen (counter-factual events) or because they are reported with uncertainty. In order to annotate only events potentially candidates to participate to a timeline, we have defined criteria based on the intra-document annotation Guidelines.

We annotate verbs, except if they are modified by a modal word, nouns and pronouns. Adjectives generally express a property or attribute of an entity, and anchoring them in time is not simple. So adjectival events will not be annotated.

Events are classified according to semantic features. Those classified as “grammatical” are dependent on a content verb/noun and don’t have a time span, so they will not be annotated. We have also decided to leave out cognitive events (i.e. events that describe mental states or mental acts).

The last criteria is based on the factuality and certainty of events. Counterfactual events will not be part of a timeline because they did not take place. Non-factual events are speculative events, so we don't know if they happen or not. If it's certain that they will happen (e.g. "the conference will take place on Monday"), they will be annotated. But if they are uncertain (e.g. "the conference may take place later"), we will not annotate them.

Following are the rules applied to select events which are potentially candidates of a timeline. Into brackets we specify the section in NWR-2014-2-2 where annotators can find more details. For each criteria we provide some examples, in which selected events are in bold.

- Part-of-speech (section 5.2.2 of NWR-2014-2-2)
 - verb (except if it is modified by a modal word)
*Apple has **claimed** that the 3G iPhone will **provide** internet access at double the speed of the internet access **provided** by the previous versions.*
 - noun
*Apple Inc. today has introduced the much-anticipated iPhone at the Macworld **Conference** in San Francisco.*
 - pronoun
*Apple Corps claims that Apple Computer's iTunes Music Store violates an agreement reached between the two companies in 1991, **which** barred Apple Computer from using the "Apple" brand.*
- Event Classes (section 4.2 of NWR-2014-2-2)
 - SPEECH_COGNITIVE: only events that describe the action of a person or an organization declaring something, narrating an event, informing about an event
*Apple Inc. **announced** Friday that it will give owners of its new iPhone 4 a free case in response to mounting concerns over the device's antenna placement.*
*Jobs also **said** that the phones can be returned for a refund as well.*
 - OTHER
*Apple Computer today **launched** a 3G version of its iPhone device.*
*He **plans** a meeting on Friday.*
- Attribution - factuality (section 5.2.3 and 5.2.4 of NWR-2014-2-2)

- factual events (certainty=CERTAIN, time=NON_FUTURE, polarity=pos)

*Apple Corps **claims** that Apple Computer's iTunes Music Store violates an **agreement** reached between the two companies in 1991, which barred Apple Computer from using the "Apple" brand in certain uses in the music business.*
- non-factual and certain events (certainty=CERTAIN, time=FUTURE, polarity=pos) (except if they are part of a conditional construction)

*Apple has claimed that the 3G iPhone will **provide** internet access at double the speed of the internet access provided by the previous versions.*

*The former Chief Operating Officer, Tim Cook will be **succeeding** Jobs as CEO.*

Following we define criteria to leave some events out (in the examples events not to be annotated are in underlined).

- POS

- adjective

The new phone will also come pre-loaded with the new iPhone 3.0 software, which will be available on June 17.

The iPod nano is now smaller and squarer.
- prepositional events

They had 150 passengers on board.

- Event Classes

- SPEECH_COGNITIVE: events that describe mental states and mental acts that involve mental or cognitive processes

He plans to go to Roma.

"We love the Beatles, and it has been painful being at odds with them over these trademarks", said Apple Inc. CEO Steve Jobs.
- GRAMMATICAL

The announcement came in a rare press conference held by Apple on Friday morning.

- Attribution - factuality

- non-factual and uncertain events (certainty=POSSIBLE/PROBABLE, time=FUTURE/NON_FUTURE, polarity=pos)

Apple CEO Steve Jobs believes that this may have a large impact on the film industry.

- counter-factual events (polarity=neg)

Papermaster did not comment on the situation.

- Specific cases:

- conditional constructions

The iPhone 4 was plagued by highly publicized reports that abnormalities in its new antenna design caused the device to lose its cellular signal if held in a certain way.

The iPhone 4 antenna is actually a metal strip that wraps around the side of the device, which has caused dropped calls when held in a certain way.

- verbs modified by modal verbs (note: according to the NewsReader guidelines “will” is considered as an auxiliary and not as a modal verb)

Any members of the iPhone Developer Program can add applications to the store for free.

- conjunction of events:

*Developers will be able to **set** the price of the applications or **re-lease** them for free. → “set” and “release” are annotated but not the discontinuous event mention “set (or) release”*

- causality (section 10.3 of NWR-2014-2-2): ENABLE-type verbs (e.g. enable, aid, allow, permit) and verbs in their complements
Features new to the iPhone include an upgraded camera, which also allows users to record video and sound, as well as ” voice control”, which will allow users to control most features of the iPhone with their voice.

3.2 Event instances

CROMER event instances have the following attributes:

- class (compulsory): to be filled according to the guidelines provided for the intra-document annotation (see section 4.2 of NWR-2014-2-2);
- name (compulsory): to be filled according to the guidelines provided for the intra-document annotation (see section 4.1 of NWR-2014-2-2 on Tag Descriptor for Events);

- short description (compulsory): a short description of the event instance, whose aim is to distinguish it from other event instances with similar names;
- time (compulsory for punctual events): a date (maximum granularity day), following the TIMEX3 format (see Section 6.2.2 of NWR-2014-2-2); if not known, put the values XXXX-XX-XX, XXXX-XX or XXXX depending on granularity;
- begin (compulsory for durative events): the starting date of the event (maximum granularity day), following the TIMEX3 format (see Section 6.2.2 of NWR-2014-2-2)); if not known, put the values XXXX-XX-XX, XXXX-XX or XXXX depending on granularity;
- end (if known, compulsory for durative events): the ending date (maximum granularity day), following the TIMEX3 format (see Section 6.2.2 of NWR-2014-2-2);
- external reference (compulsory): to be filled according to the guidelines provided for the intra-document annotation (see section 4.3 of NWR-2014-2-2);
- comment (optional).

In case of repeated events or grouped events the following rules should be applied to fill the time attributes:

- repeated events (e.g. "I go to work every morning")
 - time: if known put the value of the set temporal expression (following the TIMEX3 format);
 - begin (compulsory for repeated events): put the date of the first time the repeated event happened; if not known, put the values XXXX-XX-XX, XXXX-XX or XXXX depending on granularity;
 - end: if known, put the date of the last time last the repeated event happened;
- grouped events (e.g. "the explosions caused a hedge damages") are considered as if they were durations
 - begin (compulsory for grouped events): put the date of the event that happened first chronologically; if not known, put the values XXXX-XX-XX, XXXX-XX or XXXX depending on granularity;
 - end: if known, put the date of the event that happened last chronologically;

4 Annotation task

The cross-document annotation will be done using CROMER, a cross-document annotation tool ([Girardi *et al.*, 2014]).

4.1 Phase 1: Entity annotation

For each entity annotated in CAT in the 6 first sentences the annotation task consists of the following steps:

- check if an entity instance already exists in CROMER and if not create it;
- assign the mention chains to the entity instance; (if annotators find annotation errors in the import from CAT, they can correct them).
- find other mentions of the entity in the corpus (only in the first 6 sentences of each document annotated with CAT) and assign them to the entity instance.

Each partner institution receives a pre-defined list of entities to be annotated for each corpus.

For each selected entity the annotation task consists of the following steps:

- check if an entity instance already exists in CROMER and if not create it;
- assign the mention chains imported from CAT (which refer only to the first 6 sentences of a document) to the entity instance; (if annotators find annotation errors in the import from CAT, they can correct them).
- find all other mentions of the entity in the corpus (i.e. in the remaining part of each document), annotate them with the correct extent (no attributes have to be annotated) and assign them to the entity instance.

4.2 Phase II: Event annotation

Annotators should annotate all events having as participant one of the seed entities annotated in phase I.

More specifically, the annotation task consists of the following steps:

- For each selected entity, get the list of all documents in which that entity is mentioned.

- For each document, identify all event mentions having the annotated entity as participant and annotate them as follows:
 - check if the event instance to which it refers already exists; if it does not exist, create it;
 - create a HAS-PARTICIPANT relation between the event instance (source) and the entity instance (target);
 - if the event mention is one imported from CAT, assign it to the event instance it refers to; otherwise annotate the mention with the correct extent (no attributes have to be annotated) and assign it to the event instance it refers to.

References

- [Bach, 1986] Emmon Bach. The algebra of events. *Linguistics and Philosophy*, 9:5–16, 1986.
- [Girardi *et al.*, 2014] Christian Girardi, Manuela Speranza, Rachele Sprugnoli, and Sara Tonelli. Cromer: A tool for cross-document event and entity coreference. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*, Reykjavik, Iceland, May 26-31 2014.
- [ISO TimeML Working Group, 2008] 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>.
- [Tonelli *et al.*, NWR2014-2-2] Sara Tonelli, Rachele Sprugnoli, Manuela Speranza, and Anne-Lyse Minard. NewsReader Guidelines for Annotation at Document Level. Technical report, Fondazione Bruno Kessler, NWR2014-2-2.