Annotation Guidelines for WikiBio corpus

This document describes the annotation guidelines for the entity-based event extraction task. Given a text, which can be a sentence, a paragraph or a full document, annotators must first identify all the events related to a target entity and then:

1. Label the single word triggering the event
2. Label the word or group of words triggering the entities

As it can be observed in the example (a), only events that are directly related to the entity must be annotated. Thereby, the pair <Woods, BORN> is a proper annotation, while the pair <his family, LIVE> is not, since the target-entity has not a direct role in the event LIVE.

(a) Woods **[TARGET-ENTITY]** was born **[EVENT]** at Hobeni , Transkei , where his family had lived for five generations.

The corpus

The corpus of documents to be annotated is a set of 20 Wikipedia biographies of writers born in Africa or African American writers. The complete list of writers in the following:

Donald Woods, Ada Aharoni, Lewis Nkosi, Ngũgĩ wa Thiong'o, Angela Davis, Nasr Hamid Abu Zayd, Wole Soyinka, Bessie Amelia Emery Head, Ken Saro-Wiwa, Ali Al'amin Mazrui,Abdel-Tawab Youssef Ahmed Youssef,Alice Malsenior Tallulah - Kate Walker, Jayne Cortez, Chloe Anthony Wofford Morrison, Etheridge Knight, Maya Angelou, Amiri Baraka, John Edgar Wideman, Dwight D. York, Ishmael Scott Reed

Entities

Our guidelines for the annotation of target entities start from the Co-reference Guidelines for English Ontonotes[1], introducing some simplifications and variations.

Simplications

1. Instead of annotating the mentions of all the entities, annotators are asked to label only mentions about the target entity, namely the subject of the biography.
2. Annotators must not annotate appositive coreference as it can be observed in example (b).

[1] https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-coreference-guidelines.pdf

(b) Luke **[TARGET-ENTITY]**, ~~a writer from Somalia~~, was born in 1973

Variations

1. **Mentions without a role**. Mentions must be annotated only if they result in a direct role of the target entity in the event. SO events in which the target entity is mentioned for its relation with another entity who has a role must not be annotated, as it can be observed in (c)

(c) ~~His~~ father was the Chief Kadhi of Kenya,

2. **Metonymical mentions in biographical events.** There are some cases in which events refer to a target entity without directly mentioning it, as for instance in cases where the book of an author is awarded, translated or published. Such a relation may be considered as metonymic, since the entity that receives a prize is the author and not the book. Traditional coreference resolution guidelines ask annotators to consider such cases as mentions of the book, though. Instead in our guidelines annotators must consider these as metonymic mentions of the target-entity whenever the event is biographical, as in (d). If however the mention of the book is not related to a biographical event as in (e), annotators must not annotate it.

(d) In 1975 , Morrison's second novel Sula **[TARGET-ENTITY]** (1973) , about a friendship between two black women , was nominated **[EVENT]** for the National Book Award .
(e) The ~~book~~ talks about the relationship between a journalist and his dog.

3. "**Part of" mentions**. A last variation from the OntoNotes coreference guidelines refers to the "part of" relation between the target-entity and a group it is part of. Unlike traditional coreference guidelines, ours require annotators to label such mentions. As it can be observed in (f), the pronoun "they" is marked as a mention of the entity target, since it is involved in the event. However, whenever it is possible to distinguish the target entity, annotators must annotate only it, as in (g), where "and his wife" was not marked. Such a type of mention is also applied to groups (h), but not to organizations.

(f) He **[TARGET-ENTITY]** exhibited with his wife and they **[TARGET-ENTITY]** received enthusiastic reaction
(g) He **[TARGET-ENTITY]** and his wife won the Nobel Prize.
(h) He **[TARGET-ENTITY]** founded Nirvana. The group **[TARGET-ENTITY]** toured Europe in 1992.
(j) He **[TARGET-ENTITY]** founded Apple. ~~Apple~~ launched IPhone in 2007.

Event

Our definition of events derives from TimeML.
“We consider “events” a cover term for situations that happen or occur. Events can be punctual (1-2) or last for a period of time (3-4). We also consider as events those predicates describing states or circumstances in which something obtains or holds true (5).”
Annotators must think of events as something that occurred in a certain moment or for a certain period of time within the life of the entity.

In order to correctly detect events there are four aspects you must pay attention:

1. One event == one token

When you are annotating an event you must try to always annotate only one token. This means that you must not annotate auxiliaries (a), prepositions for phrasal verbs (b), and other words that form a MWE (c).

a. He ~~has been~~ awarded
b. He grew ~~up~~ in Ogidi
c. He is the ~~US~~ President since 1999

2. Events may be expressed by several part of speech

Even if they are more frequently expressed by verbs, EVENTS may be also expressed by other parts of speech, such as names, adjectives, and pronouns. In this task annotators must annotate events regardless of their part of speech, as it can be observed in (d), (e), and (f).

d. He won **[EVENT]** the Nobel Prize
e. He has been professor **[EVENT]**at Berkeley for 5 years
f. He was really sad **[EVENT]** yesterday

3. Light and copular verbs

Not all verbs trigger events. There are in fact several verbs that do not express events, such as copular verbs and lexical items participating in light verb constructions. These verbs are often semantically void, but may have a role in specializing the semantic of an event, for instance providing information about aspectuality. In our guidelines we ask annotators to pay attention to the following verbs that may be light or copular:

- be, become, seem, have, do, make, get, come, put

If it is so, annotators must label them as REL (Bonial, Palmer, 2016) and link them to the event they refer to, as it can be observed in (g) and (h). However, these verbs may also express an event alone, asi in (i) and (j).

g. He make **[REL**] a speech **[EVENT]**
h. He get **[REL]** a scholarship **[EVENT]**
i. They were **[EVENT]** in Greece for 6 weeks
j. He made **[EVENT]** a cake.

4. Annotating uncertainty

Uncertainty is a crucial aspect in annotating events, since it may affect time reasoning, and it is crucial to the domain we are investigating. If a person “tries to be elected in Parliament”, it is important to label the event “elected”, but at the same time to mark the uncertainty of such an event. Annotators are asked to:
1. identify in the text events or other linguistic items that express uncertainty,
2. label them as EVENT if they are events or EVENT_MOD if they are not.
3. link them to the event that they are related to.

There are three types of uncertainty links:
- INTENTION: if the event represents the intention of an agent
- NOT_HAPPENED: the event didn’t happen
- EPISTEMIC: all the other cases. In particular events related to opinions and hypothetical events.

k. the government was trying **[EVENT]** to have him killed **[EVENT]**. <trying, killed, INTENTION>
l. was not [EVENT-MOD] allowed to speak [EVENT] publicly. <not, speak, NOT_HAPPENED>
m. Dr Mamphela Ramphele , berated **[EVENT]** him for writing **[EVENT]** misleading stories about the movement <berated, writing, EPISTEMIC>