Guidelines for the annotation of full texts

formerly known as "guidelines - coref-annotation"

last modification: 2014-02-25

Summary

This document describes the annotation guidelines followed for the second round of annotation of the "CockrACE" corpus; it is derived from the annotation guidelines for the original CockrACE corpus.

Peculiarities

- We assume the annotator works with the Recon tool, which supports the annotation of
    - entity mentions of arbitrary type, plus the attachment of arbitrary properties to these entity mentions
    - relations of arbitrary type between an arbitrary number of entity mentions
- The annotator revises the original CockrACE v1 annotation, ie. we state when certain annotations have to be transformed or should be removed. Note struck-through text.

Overview of entity and relation types

Entities

- Individual entities (person, location, organization, date)
- Groups of entities (person_group)
- Heads of entities
- ~~Events~~
- Semantic terms

Relations

- Coreference chains (identity, ~~inferred~~)
- Entity_head relations
- Member_of_group relations
- Semantic relations (marriage, person_parent, sibling_relationship)

1. Entity Types

1.1. Individual entities

Individual entities may be of one of the following types: person, location, organization, date. For the rest of this section, we use person entities as an example, but locations, organizations, and dates are annotated in the same fashion.

Person entities may be referred to by their name, a noun or nominal phrase, or by a pronoun. Thus a single person entity can be expressed in several ways:

entity	*Jon Bon Jovi*
name	Jon Bon Jovi, Mr. Bon Jovi, Jon,
nominal	the rocker, a member of that society, her father, the 40-year-old
pronoun	he, him, I

Each mention that represents a distinct person in the real world is annotated as type person. People may often be alternatively referred to by their name, occupation, family relation, pronoun, etc. or by some combinations of these.

In particular, we annotate as type person: name mentions, nominal mentions, pronoun mentions, religious figures, fictional characters, positions.

1.2. Groups

Groups of people are to be considered an entity of type person_group unless the group meets the requirements of an organization, which we state as a formal organizational structure within the group.

We annotate as type person_group:

name mentions	the Bon Jovis
nominal mentions	the couple, the newlyweds, the three men
pronoun mentions	they, their

Counter-example: *George and Barbara Bush*

1.3. Events

The formerly annotated event entities, which were represented by their name, a noun or nominal phrase, by a pronoun, a verb or by an adjective phrase, are not used anymore. Instead, we delete them, or we change them to entity type semantic_term and change their role in relation mentions to related.

1.4. Semantic Terms

Every word which semantically expresses the core idea of our target relations or which is closely related to a semantic relation is to be marked and labelled as „semantic_term".

his sister	first entity	*his sister*
	second entity	*his*
	semantic term	*sister*
their marriage	first entity	*their*
	semantic term	*marriage*

Note the "term" in "semantic terms", i.e. there is a focus on individual words or short compound nouns. Considerations about the extent of mentions and their head (see next section) do not apply.

1.5. Entity Heads & Notes about the Extent of Mentions

➤ Definite and indefinite articles are never part of an entity mention unless they are part of the name.

Example with "the" being a part of the entity mention: *"[The Beach Boys] are an American rock band …"*

➤ Wh-words are never part of an entity mention. Examples:

…[addicts] who seek treatment…
…[opponents of the proposition], many of whom are in law enforcement

Watch for elliptical constructions of the type:

…[kidnappers] trapped in a small room with him…

...the [man] himself hanging out of helicopter in Proof…

➤ <u>Head</u> entities may be represented by a name (*Laura Bosley*), noun (for ex., "*daughter*") or nominal phrase (for ex., terms or concepts: "*maid of honour*"). Head entities are annotated as separate entities, but do never occur on their own; they are always connected to an entity mention of type person, person_group, location, organization, or date via a mention of the relation <u>entity_head</u>.

Example: "*FBI spokeswoman Laura Bosley*" has as its head "*Laura Bosley*", ie. the annotation should be as follows:

[*FBI spokeswoman* [*Laura Bosley* | <u>head</u>] | <u>person</u>]

2. Co-reference Annotation: Identity Relations

Referring expressions point back to the same entity.

Every mention of type <u>person</u> or <u>person_group</u> which refers to the same entity in the real world and which follows each other within the text structure should be linked by an <u>identity</u> relation. Thus we build a co-reference set for every single <u>person</u> or <u>person_group</u> entity mentioned in a text. We do not establish more than one set per <u>person</u> or <u>person_group</u>, so that the sets include every single possible mention of that <u>person</u> or <u>person_group</u>.

Examples:

<u>identity</u> <*her husband*, *British film director*, *Guy*, *Guy Ritchie*, *Mr. Ritchie*, *dad*, …>

<u>identity</u> <*they*, *the couple*, *the newlyweds*, *the parents*, …>

Coreference sets only consist of mentions that point to the exact same entity.

2.1. Source and target

Any type of co-reference relation consists of two parts, the "<u>source</u>" and the "<u>target</u>". While the <u>target</u>, which may be an anaphoric or cataphoric mention, is somehow related and referring to an antecedent within the text structure, the <u>source</u> is a name, noun or nominal phrase which directly represents the entity of the real world. With these two labels anaphoric and cataphoric relationships between two mentions can be shown.

Anaphoric expression:

Guy Ritchie (<u>source</u>) *followed, turning guest's heads as he* (<u>target</u>) *entered in a vibrant teal blazer...*

Cataphoric expression:

Madonna's 4-year-old-daughter (<u>target</u>)*, Lourdes* (<u>source</u>)*, ...*

2.2. Inferred relations

<u>Inferred</u> relations were formerly used to mark a weaker form of co-reference than the <u>identity</u>

relations denote. They are not used any longer. Existing annotations should either be changed to an identity relation, be modified to be a member_of_group relation, or should simply be deleted.

3. Member_of_group relation

Whenever a group of people is introduced and (some of) its members are named or counted, we build a member_of_group relation between all mentioned parts of the group and the group itself. We do this only once per group and document right with the first mention or beginning of the mention chain. If we can not find any name mention for the members, we follow the NearFirst rule (see 6.1.) and add nominal or pronoun mentions to the existing relation.

They were strangers on a plane, three men who probably were thinking of the business meetings that awaited [*them*|person_group] *in California. It was a fluke that* [*Jeremy Glick*|person]*,* [*Tom Burnett Jr.*|person] *and* [*Mark Bingham*|person] *were all on the Boeing 757.*

- member_of_group: <group={them}, member={Jeremy Glick, Tom Burnett Jr., Mark Bingham}>
- Note: This example paragraph has only been annotated with the entity mentions relevant for this relation instance.

4. Entity_head relations

Sometimes one or more of the entities of a relation is in the form of a noun phrase, which is often ambiguous on the surface. For the sake of clarity and thoroughness we add this new relation, so we can avoid overlapping entities. We put the two related entities into the arguments head and full_extent, respectively, with the last one being the whole syntagm:

[*Madonna's 4-year-old* [*daughter* | head] | person]

entity_head <head={daughter}, full_extent={Madonna's 4-year-old daughter}>

5. Semantic Target Relations

5.1. marriage

Size constraint: At least one entity of type person or person_group must be entered.

Gender-specific arguments: Choose most specific one. Do not enter an entity for more than one argument in the same relation mention.

argument	type
person (gender unknown)	person / person_group
wife	person
husband	person

related	semantic term
from	date
to	date
ceremony	location

5.2. sibling_relationship

Size constraint: At least one entity of type person or person_group must be entered.

Gender-specific arguments: Choose most specific one. Do not enter an entity for more than one argument in the same relation mention.

argument	type
person (gender unknown)	person/person_group
sister	person/person_group
brother	person/person_group
related	semantic term

5.3. person_parent

Size constraint: At least one entity of type person or person_group must be entered.

Gender-specific arguments: Choose most specific one. Do not enter an entity for more than one argument in the same relation mention.

argument	type
person (gender unknown)	person/person_group
son	person/person_group
daughter	person/person_group
parent	person/person_group
mother	person
father	person
related	semantic term

6. Further Annotation Hints

These are some basic rules we set for the annotation process:

5.1. NameFirst vs. NearFirst

Coreference links, member_of_group:

When annotating simple co-reference links between groups and persons as well as member_of_group relations, one should normally look for a name mention to attach the current item to ("NameFirst"). Only if there is no name mention or if the name mention cannot be chosen for a reason, the nearest mention is used ("NearFirst").

Semantic relations: marriage, sibling_relationship, person_parent

When there are several mentions of an entity that could potentially be added as an argument to a relation mention, the entity mention which occurs closest to the trigger word of the relation mention should be used ("NearFirst").

5.2. Examples

➤ We have often appositions. Let us take a look at some examples:

1) *...Robert's wife, Jane...*
2) *...his wife, Jane…*
3) *...his wife Jane*
4) *...his wife...*
5) *On Ritchie's side were his mother, Lady (Amber) Leighton, and his father, John, accompanied by stepmother Shireen Ritchie...*

⇒ We always annotate as part of a relation of type person_parent, sibling_relationship, or marriage the person mention beyond the comma. Our examples would look as follows:

1) *...Robert's wife, Jane...*
marriage <<u>husband</u>: *Robert*, <u>wife</u>: *Jane*, <u>related</u>: *wife*>

- ! Do not forget the relation entity_head!
- Examples 2) and 3) are to be treated exactly the same way as 1)!

4) *...his wife...*
marriage <<u>husband</u>: *his*, <u>wife</u>: *his wife*, <u>related</u>: *wife*>

- ! Do not forget the relation entity_head!

5) *On Ritchie's side were his mother, Lady (Amber) Leighton, and his father, John, accompanied by stepmother Shireen Ritchie…*

person_parent <<u>son</u>: the first *his,* <u>mother</u>: *{Lady (Amber) Leighton, Shireen Ritchie},* <u>father</u>: *John*, <u>related</u>: {*mother, father, stepmother*}>

⇒ Some remarks on the relation <u>entity_head</u>: Despite *his wife Jane* or *Robert's wife Jane* being an apposition we always annotate them as one entity (person) with head *Jane*.
Another example: *Giuliani's divorce lawyer Raoul Felder*

- entity_head <u>full_extent</u>: *Giuliani's divorce lawyer Raoul Felder,* <u>head</u>: *Raoul Felder*>

➤ Embedded relations:

...Chynna Phillips, daughter of onetime Mamas and the Papas stars John and Michelle Phillips…

person_parent <
<u>daughter</u>: *Chynna Phillips*
<u>parent</u>: the person group *onetime Mamas and the Papas stars John and Michelle Phillips*
<u>related</u>: *daughter*
>

entity_head <
<u>full_extent</u>: *onetime Mamas and the Papas stars John and Michelle Phillips*
<u>head</u>: the head *John and Michelle Phillips*
>

member_of_group <
<u>group</u>: *onetime Mamas and the Papas stars John and Michelle Phillips*
<u>member</u>: the persons *John* and *Michelle Phillips*
>

Another example for 'embedded relations':

...12-year-old twins Amanda and Jonathan Schwebel...

sibling_relationship <
<u>sister:</u> *Amanda*
<u>brother:</u> *Jonathan Schwebel*
<u>related:</u> *twins*
>

entity_head <
<u>full_extent</u>: *12-year-old twins Amanda and Jonathan Schwebel*
<u>head</u>: the head *Amanda and Jonathan Schwebel*
>

member_of_group <
<u>group</u>: *12-year-old twins Amanda and Jonathan Schwebel*
<u>member</u>: the persons *Amanda* and *Jonathan Schwebel*
>

➤ We annotate reasonably long markables, i.e. we focus on the mention heads.

... [[a member] of that small society of aging-but-blazing [singers] who've learned how to stoke fans while keeping the home fires burning] ...