Annotation Guideline

Berfin Aktaş and Manfred Stede

08 Oct 2019
Version 2.1

1 Introduction

This guideline presents the instructions for the annotation of coreference chains in conversational social media texts[1]. This version of the guideline addresses only identity coreference; non-identity reference (bridging) is not being annotated for now. For the annotation, we use freely available MMAX annotation tool[2].

In the following, Section 2 describes in detail the types of referring expressions that are subject to the annotation. Section 3 describes the annotation process, and Section 4 defines the attributes that have to be assigned to each markable.

2 Markables

In this section, we first discuss the various types of markables to be annotated in 2.1, and then in 2.2 provide guidance on identifying their spans. The square brackets around the expressions demonstrate the span boundaries and indices under the brackets represent the id of the coreference chain where the markable belongs to.

2.1 Types of markables

Syntactically, markables are phrases with nominal or pronominal heads. The following referring expressions are to be considered as markables:

1. Full nominal phrases, e.g. *a big blue sky*;
2. Proper names and titles, e.g. *Mr. Black*;
3. Pronouns
   - Personal pronouns: We annotate the personal pronouns "*I, you, my, me, mine, your, yours, they, them, their, theirs, he, him, his, himself, her, hers, her, herself, it, its, itself*".

     (1) [**It**]$_j$'s beginning to rain! - [Daisy]$_i$ exclaimed to [**herself**]$_i$.
     (2) [John]$_i$ is calling [**his**]$_i$ doctor.
     (3) [**She**]$_i$ is [my new lawyer]$_i$.
     (4) [**I**]$_j$ have already payed 699 for [this]$_i$ and [**it**]$_i$ is not working.
     (5) This is [**my**]$_i$ notebook. [**I**]$_i$ bought [**it**]$_j$ last week.

     In multi-turn conversations with more than two participants, it is possible that first, second and third person pronouns refer to the same entity.

     (6) **user1:** [I]$_i$ prefer to go to small cinemas instead of the big chains.
     (7) **user2:** (reply-to-user1) What are [you]$_i$ talking about?
     (8) **user3:** (reply-to-user2) [He]$_i$ is talking about supporting small business.
   - Demonstrative pronouns: *this, that, these, those*

[1] Most relevant portions of "Parellel coreference annotation guidelines" by Yulia Grishina and Manfred Stede are adapted to our data.

[2] http://mmax2.net/index.html

(9) You need [a camera that works in the dark]$_i$. Hm, take [**this**]$_i$, [it]$_i$ has a great shutter speed.

In the example, the demonstrative pronoun *this* corefers with the pronoun *it* in the next sentence and must be annotated.

Predicative constructions are annotated in the following way:

(10) [This]$_i$ is [a bank]$_i$, but [it]$_i$ is not very well-known.

- Relative pronouns, such as *who, whom, whose, which, that* etc.

  The relative pronouns are used to construct relative clauses in the sentence. There are different types of relative clauses for which the annotation instructions are presented below.

  If a relative pronoun is used in a restrictive relative clause, the whole NP span is annotated as one mention:

  (11) I met [the cyclist who won the race]$_i$. [She]$_i$ deserved that result.

  We use non-defining relative clauses to give extra information about the person or thing. In writing, commas are often put around non-defining relative clauses. In that case the modified noun and the relative pronoun are annotated separately and put in the same coreference chain.

  (12) I was talking about [my uncle]$_i$, [who]$_i$ has the horse, when you came.

  There exist also free (headless) relative clauses which are not used as noun modifiers, instead they serve as arguments in the main clause. In that case, we only annotate the relative pronoun.

  (13) I saw [what]$_i$ you cooked and ate [it]$_i$.

  Keep in mind that pronouns can be ambiguous:

  (14) For both India and Pakistan, Afghanistan risks turning into a new disputed territory, like [Kashmir]$_i$, [where]$_i$ the conflict has damaged both countries for more than 50 years.

  (15) Daisy managed to discover *where* Mr. Baccini's dishonest partner was now living and was anxiously expecting her cheque.

  In example 14, "where" is a relative pronoun and refers to Kashmir (to confirm this, one can substitute *where* by *in which*). In contrast, in 15, *where* is not a relative pronoun and should not be annotated.

- Question pronouns, such as *who, what, which, where* etc.

  Question pronouns are not considered as markables and not annotated.

  For instance, we don't annotate "who" in the following small conversation:

  (16) – **user1:** Who is calling?
  – **user2:** [Jane]$_i$ is on the phone, I think [she]$_i$ wants to visit us.

- "HIS/HERS", "HIS or HERS" and similar forms are annotated as a single markable.

4. NPs with quantifiers

   Be careful when annotating NPs with quantifiers, e.g. *all people, two people, 105 Million euro* etc. If you are not sure about the definiteness of an NP, apply the following test: try inserting a definite article or a demonstrative pronoun. If the meaning of the phrase is not changed, then the NP is definite. Example: "*all people*" > "*all these people*" > definite NP.

   Quantifiers (of the form "X of Y") should not be coreferenced with the entities they modify:

   (17) "[a mile of [highway]$_j$]$_i$"

   (18) "[a group of [doctors]$_k$]$_l$"

   Similar constructions such as "both of those things" and "all of my friends", the markable spans are similar but coreferential strategy is different: Since "my friends" and "all of my friends" would usually be equivalent/refer to the same group of people, both ("my friends" and the full phrase "all of my friends") are selected as markables and they are marked as coreferent.

5. Nominal premodifiers

   Nominal premodifiers are annotated as separate markables only if there is an overt reference to that modifier in the text as in 19. Otherwise they are not annotated separately and included in the span of the NP they modify (20).

   (19) I bought a new [[ceramic]$_i$ pot]$_j$. I really like [that material]$_i$ for cooking because nothing sticks on [it]$_i$.

   (20) I bought a new [ceramic pot]$_i$. [It]$_i$ is perfect for frying!

6. Groups

   Antecedents of plural pronouns can be non-contiguous. In that case, we follow the strategy explained below through example 21.

   In example 21, *your husband* and *Mrs. Humphries* constitute the antecedent for the plural pronoun *we.* But as they are non-adjacent markables, we can't annotate them as a group. Therefore, we annotate them as markables but do not corefer with the pronoun *we.*

   (21) Did [[your]$_i$ husband]$_j$ buy Lorna, [Mrs. Humphries]$_i$? - No, [we]$_k$ bought her together.

7. Coordinated NPs

   Coordinated NPs are annotated both separately and as a whole

   (22) [[lies] and [assumptions]]

8. Numbers are annotated only if they are nominalized.

   (23) There were 100 participants in the meeting. [5]$_i$ among them was selected for the next step. [It]$_i$ was [an ambitious group]$_i$.

   (24) [The first]$_i$ is [a woman]$_i$. [She]$_i$ will join us later.

9. Temporal Expressions are annotated.

   (25) We are going to meet on [Thursday]$_i$ because [it]$_i$ is [Anna's birthday]$_i$.

   The deictic temporal expressions (e.g. today, tomorrow) are also annotated with the following values assigned to relevant fields:

   - TODO: should be specified here.

10. Do not annotate non-referential NPs in idioms, or lexicalized phrases such as "for example","A penny for your thoughts" etc..

11. Predicative forms

    In simple copula relations, the mentions corefer. When a copula relation is negated, the mentions should not corefer.

    (26) (a) [Oxford]$_i$ is [a university]$_i$. [It]$_i$ has a long history.
    (b) [John]$_i$ is not [a lawyer]$_i$, [he]$_i$ is [an architect]$_i$.

12. Non-nominal referents (e.g. clauses, propositions, verbs) are not annotated. But the pronouns and nominal expressions referring these non-nominal antecedents are annotated and "referent_type" for these expressions should be selected as "clausal".

    (27) There is a big growth in the economy in last year. [This] is very surprising in current conditions.

    The antecedent for *This* is the whole sentence "There is a big growth in economy in last years.". In our annotation scheme, we do not annotate this sentential antecedent but we annotate "This" and assign the referent type "clausal" to this markable.

13. *One* pronoun

    "One" is annotated as a pronoun in the cases similar to the following:

(28) "We have [one] (ellipsis: calendar) up for grabs.

14. Usernames and hashtags

    In Twitter, there are automatically inserted usernames in the replies and also hashtags added by the users to increase the visibility of the tweets. The automatically inserted usernames (at the beginning of the tweet) and hashtags are not annotated unless they are referred by other expressions. The special cases in which the usernames and hashtags should be annotated are exemplified below:

    (29) [@BarackObama]$_i$ should change [his]$_i$ policy.

    (30) Yes! [She]$_i$ is my favorite. [#Oprah]$_i$

15. Emojis

    Emojis are annotated if they are used in place of nominals as in the examples below:

    (31) He really loves [that 🐍].

    (32) Can you please drop by [our ⭐ baker]?

We assign specific values to the *correct_form* and *comment* attributes in the scheme if the mention span contains an emoji:

- the field "comment" contains the string "emoji"
- The field "correct form" is for the string representation (e.g. "that snake" for the emoji-containing mention in 31 and "our star baker" for 32).

2.2 Spans of markables

Markables are always rooted in some nominal phrase (NP), and their extension is defined as follows:

- The syntactic head of the NP;
- Determiners and adjectives (if any) that modify the NP;
- Dependent prepositional phrases (for example, [Queen of England]).

Appositions, i.e., additive material that is not syntactically integrated are annotated separately (check example 51 for this case). Nested mentions can also exist in the text, we allow the annotation of nested markable spans:

(33) In our colloquium today, [Slavoj Žižek]$_i$ will be talking about [[his]$_i$ new book]$_k$. [It]$_k$ is published by MIT Press.

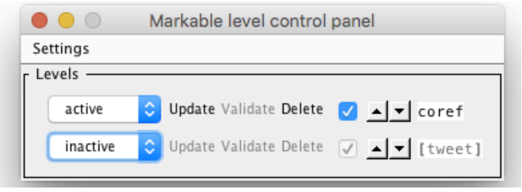
We do not allow discontinuous markables. Punctuations such as comma, paranthesis, question mark etc. are not involved in the markable span if they are not part of the proper name:

(34) He started to work in [Yahoo!].

3 Annotation Process

The annotation process selects all nominal expressions (‘markables’). We annotate the complete reference chains for each entity referred by a nominal expression. Therefore, the annotation process involves a certain amount of “going back and forth” in the text.

If there are other annotation levels active in the scheme, the levels other than the "coref" level should be set to "inactive" to make the annotation process simpler as in the figure below:



Markable level control panel

Settings

Levels

active Update Validate Delete ☑ coref

inactive Update Validate Delete ☑ [tweet]

Annotation of coreference chains is an incremental process. First step is to highlight the referential nominal expressions (i.e., names, noun phrases and pronouns) in the text. This is done by selecting the "Create Markable on level 'coref'" option in MMAX. After all the mentions are highlighted in "coref" level, the ones referring to the same entity should be put in the same coreference chain. Each mention should be linked to the closest antecedent by selecting the "Mark as coreferent" option in MMAX. In case of cataphoric pronouns ('Before she left, Sue locked the door') the relation is to be established in forward direction (here: from 'she' to 'Sue'). The exophoric pronouns which refers to an entity out of the text (i.e. there is no overt antecedent in the linguistic domain) should also be annotated. In some cases, these exophoric pronouns may create singleton chains if there is no other mention in the text referring to the same entity as the exophoric pronoun. Non-referential (pleonastic) pronouns should also be annotated as markables and they will be considered as singleton coreference chains as no other mention is linked to them.

4 Attributes

4.1 Attributes for all markables

4.1.1 representative_men

With this attribute, we identify the most informative/descriptive mention as the representative mention of the entity. In general, here is the hierarchy in terms of representativeness among the mentions: NE>defNP>indefNP>pronoun. If there are more than one mention which can be considered as the representative of the coreference chain, select the first instance in the text. Every coreference chain should have only 1 representative mention.

4.1.2 np_form

1. none - not a nominal entity (no markable should be assigned to this type for this version of the guideline)
2. ne - named entity
3. defnp - definite NP
4. indefnp - indefinite NP
5. ppers - personal pronoun
6. ppos - possessive pronoun
7. padv - pronominal adverb
8. pds - demonstrative pronoun
9. prel - relative pronoun
10. prefl - reflexive pronoun
11. other - none of these options

Possessive pronouns mine/yours/his/hers/ours/theirs are coreferent with the possessed item, not the possessor (except for some special cases like in "friend of mine" where *mine* is coreferential with the other first person pronouns referring to the same speaker).

4.1.3 genericity

This value shows whether the referential expression under concern is referring to a specific entity or whether it is a generic nominal expression. Genericity is only assigned to the representative mention.

Please note that the plurals without a determiner, singular nouns with indefinite determiner (a/an) and NPs with "any/no" such as "no facts/no man/any person" are likely to be generic NPs. For all the NPs, it requires individual investigation of the NPs. There are also generic usage of pronouns as exemplified with the second person pronoun *you* in the examples below. Please note that the pronoun type for the generic pronouns is "exophora -> symbolic_deixis".

Nominal expressions are *generic* in the following cases:

(35) "[Parents]i_generic should take care of [their]i children."

(36) If [a man]i_generic says something like that,...

(37) Hmm, [you]i_generic can't really tell what has happened there. That incident is too complicated.

Singleton relative pronouns like this are annotated with genericity value "specific":

(38) you don't know [who] he's caring for.

4.1.4 grammatical_role

This attribute describes the grammatical role of the annotated mention or the grammatical role of the higher level nominal phrase where the annotated mention belongs to.

1. none - The mention is not part of the syntax.

   (39) We need to find [her]$_{\mathrm{i}}$. [#ClaudiaJohnson]$_{\mathrm{i}}$ (grammatical role for #ClaudiaJohnson should be chosen as *none*)

   (40) Yes! [She]$_{\mathrm{i}}$ is my favorite. [@Oprah]$_{\mathrm{i}}$ (grammatical role for @Oprah should be chosen as *none*)

2. sbj - The mention or the NP that this mention belongs to is a subject.

3. dir_obj - The mention or the NP that this mention belongs to is a direct object.

4. indir_obj - The mention or the NP that this mention belong to is an indirect object.

5. prep_phrase - The mention or the NP that this mention belongs to is a prepositional phrase.

6. copula_rel - The mention or the NP that this mention belongs to is part of the copula relation.

7. adv - The mention or the NP that this mention belongs to is part of an adverb.

8. other - none of these options

(41) Find [her]$_{\mathrm{i\_directObject}}$! [#ClaudiaJohnson]$_{\mathrm{i\_noGrammaticalRole}}$

(42) I gave [his]$_{\mathrm{i\_directObject}}$ wallet to [her]$_{\mathrm{i\_indirectObject}}$.

(43) Just check [[his] stats]: "his stats" is the "dir_obj", and the possessive pronoun before the noun is also marked as the same grammatical role ("dir_obj")

(44) Coming back to [his]$_{\mathrm{i\_prepositionalPhrase}}$ house soon.

(45) [This]$_{\mathrm{i\_subject}}$ is [[his]$_{\mathrm{j\_copulaRelation}}$ favorite]$_{\mathrm{i\_copulaRelation}}$.

(46) He is coming [today]$_{\mathrm{i\_adv}}$.

The adverbs (if they are referring mentions) *here/there* are annotated as np_form–>padv and pronoun type will be "exophora -> symbolic deixis" if they are used to refer to the locations. For other adverbs like *home/today/yesterday...* np_form is assigned to defnp even some of them are deictic (e.g. the temporal adverbs today, tomorrow, next year etc.).

The grammatical role of reflexives could be not so clear as in the case below: Reflexive pronouns used for emphasis are annotated as appositives TODO: check how this is annotated, appos?

(47) [I]$_{\mathrm{i\_subj}}$ did it [myself]$_{\mathrm{i\_appositive}}$.

So for the reflexive pronoun, if it is clear that the pronoun is the object of the sentence, or prepositional phrase, assign the relevant grammatical role to the markable. But in all the other cases (as in the example above) assign the grammatical role "other" to the markables.

(48) [He]$_{\mathrm{i\_sbj}}$ presents [himself]$_{\mathrm{i\_dirobj}}$ as a change-maker.

(49) [He]$_{\mathrm{i\_sbj}}$ cooked [himself]$_{\mathrm{i\_indirobj}}$ a delicious cake.

(50) [He]$_{\mathrm{i\_sbj}}$ prepared the food for [himself]$_{\mathrm{i\_prepphrase}}$.

other_grammatical_role

More descriptive information on the grammatical type if the grammatical_role is selected as "other".

1. appositive - The referring expression is an appositive construction that modifies an immediately-adjacent noun phrase (which may be separated by a comma, colon, dash, or parenthesis).
2. vocative - The referring expression is a direct address to one of the participants in the conversation.
3. other - none of these options

(51) I called [Till]$_{\text{i}}$, [my friend]$_{\text{i\_appositive}}$, to invite [him]$_{\text{i}}$ to join us.

(52) [United Kingdom]$_{\text{i}}$ ([UK]$_{\text{i\_appositive}}$) has the world's fifth-largest economy. [It]$_{\text{i}}$ has a high-income economy.

(53) In case of situations such as "[you] [guys]", "[you] [fucking morons]": The pronoun is annotated as usual, "guys"/"fucking morons" as **defnp** and grammatical role as "appositive" and they are marked as coreferent.

(54) Hey [@lynda]$_{\text{i\_vocative}}$, are you going to join us today?

4.1.5 semantic_class

For the sake of simplicity, this attribute is assigned only to representative mention in a coreference chain.

1. none - The mention is a non-referential pronoun.
2. abstract - The mention refers to an abstract concept.
3. human - The mention refers to a human, including fictional characters.
4. org - The mention refers to an organization.
5. loc - The mention refers to a location.
6. pyhs_obj - The mention refers to a physical object.
7. event - The mention refers to an event. (e.g. *hurricane, heart attack* etc.)
8. time - The mention refers to a certain time.
9. other - none of these options

(55) [True love]$_{\text{i\_abstract}}$ is rare, [it]$_{\text{i}}$'s [the only thing that gives life real meaning]$_{\text{i}}$.

4.1.6 pronoun_type

1. none - The interpretation of the expression does not depend on other mentions in or out of the text.
2. non-referential pronoun - These pronouns are semantically empty, and so, refers to no entity.

   For instance, *it* pronouns in the following examples should be marked as non-referential pronouns.

(56) It's raining. (weather)

(57) It takes 4 hours to go to Minneapolis. (time)

(58) It seems that John is a good football player. (usage with a raising verb "seem", e.g. appear, look, mean, happen)

(59) It is known that… (usage with a cognitive verb, e.g. think, believe, know, anticipate, recommend etc.)

(60) It is clear that we should decline...

(61) You can make it! (part of the idiom, make it=succeed, e.g. on the face of it)

3. anaphora - The expression refers to a backward phrase in the text.
4. cataphora - The expression refers to a forward phrase in the text.
5. exophora - The expression refers to extra linguistic context.
6. bridging - A definite NP picks up some aspects of a previously introduced referent and enters into a relation with that referent other than identity. (This attribute is not used in the scope of this version of the guideline but we kept it for future changes in the annotation scope.).

exophora_type

Type of the exophoric mention

1. symbolic_deixis - Pronouns point to a referent not inside the text but in the situation of utterance (e.g. spatio-temporal or speaker knowledge is required to interpret the pronoun). Usually first and second person pronouns are considered as symbolic deixis. We annotate the first occurence of these pronouns **in every chain** as exophoric deictic but the other pronouns referring to the same entity, and so belonging the same coreference chain, are marked as anaphoric.
2. antecedent_in_attached_picture - The antecedent of the pronoun is not in the linguistic context but in the visual media attached to the text.
3. antecedent_in_attached_text - The antecedent of the pronoun is not in the current linguistic context but in the text pointed by the link attached to the text.
4. antecedent_in_quoted_tweet - The antecedent of the pronoun is not in the linguistic context but in the quoted (embedded) tweet.
5. antecedent_inferred_by_world_knowledge - The antecedent of the pronoun is not in the linguistic context but can be inferred by world knowledge.

referent_type

Type of the referent (referred expression)

1. nominal - The referent of the pronoun is a nominal entity.
2. clausal - The referent of the pronoun is a clausal entity. As we only annotate nominal referential expressions, the clausal referring expressions are not annotated. Only the nominals referring to these clauses can be annotated.

   (62) John didn't call me yesterday. [This]$_i$ made me sad.

   In the example above, the pronoun *This* refers to the whole clause *John didn't call me yesterday.* But we don't annotate the clausal expressions. Therefore, *This* is annotated but its clausal antecedent is not annotated.
3. other - none of the above

4.1.7 correct_form

If there is a typo or misspelling in the surface form of the mention, the correct spelling is written here by the annotator.

4.1.8 comment

We use this attribute to add more information about the mentions. This field is available for free comments but we also use is to add some more features about the annotated mentions, if necessary. Instead of adding new attributes for the features below, we use this field to add more information on the markable:

- **metadata**: If the mention is not part of the syntax but instead part of the conversation or message structure offered by the communication media such as usernames automatically added to the replies. Please note that the "grammatical_role" should be selected as "none" if the "comment" is set to "metadata" value. **TODO: check whether we annotate these in real data!!!**

  (63) *(Tweet_ 1)* **@StarTimesKenya:** [@dennisclaude89]$_{i\_metadata}$ [you]$_i$ should downgrade your account.
  (Reply_ to_ Tweet_ 1) **@splinister:** @StarTimesKenya @dennisclaude89 I don't think [he]$_i$ needs to downgrade...

  Please note that in the example above, only the first instance of **@dennisclaude89** is annotated, the other instances which automatically added to all the replies will not be considered as markables.

- **emoji**: If the annotated mention span contains an emoji sign, the string "emoji" is assigned to comment field.

5 Interesting cases

- In the constructions similar to following examples, the name is marked as coreferent with the person/pronoun but not with the NP "name": [[My]$_i$ name]$_j$ is [Jordan Smith]$_i$.
- "bro", "dude", "dear" etc. are annotated as npform "defnp" and grammatical role "other > vocative"
- NP expressions like "such usage", although they are anaphoric, are annotated as NP and without a pronoun type.
- "I am skeptical of [adjectives being used without [their nouns]]": both are PPs, recursive/nested grammatical roles are annotated as what they are (regardless of the recursion).
- Speaker A: What [a liar]$_i$
  Speaker B: "R u calling [me]$_j$ [a liar]$_i$?
  Both "a liar" are marked as coreferent, but not coreferent with "me" (as it is contentious)
- "[Florida representative]$_i$[Matt Gaetz]$_i$": The name "Matt Gaetz" is the representative mention, "Florida representative" is marked as "other > appositive"
- "[Indian Hindu] support [Isreal]": Both are marked as "org" because they refer to political groups/governments.
- "Let ['s] talk about": "'s" (us) is annotated as "generic" and "ppers".
- "you don't know [what] [their intentions] are": "what" is annotated as "prel" and it is a singleton.
- "if [he]$_i$ said outright '[I]$_i$ don't like gay marriage'": The pronouns are coreferent, regardless of "if" and direct speech.
- "In [2018]", "In [March]" ... : Years and months are annoated as named entities (i.e. np_form *ne*).
- The relative pronoun "what" can take on many pronoun_types:
  - it may be exophora -> antecedent_inferred_by_worldknowledge if the referent is known and outside the text ("[what] Obama said"),

  - anaphoric if it refers to a already mentioned referent (nominal: "[This]i is [what]i I mean") or the antecedent can be derived from context (clausal: "[What] you said here")
  - cataphoric in sentences like "[What]i I need is [time]i" (nominal) or "[What] they talked about was how ..." (clausal)
  - If "what" introduces a new antecedent, pronoun_type might be *none* for the first mention ("I saw [what] you cooked")

- "both" is annotated as np_form "other".

- Here/There: Every occurence is annotated for consistency and as "symbolic_deixis" in the first and/or representative mention. If there is another here/there in the same chain, it is annotated as "anaphora". To differentiate between local here/there ("here in Australia") and other uses ("Yet here we are" (in this situation), "Here on Twitter"), the first is annotated as semantic_class "loc", the non-local use as "other".

6 Validation Checks

Following validations are done on the data with the automated scripts:

- All the chains (including singletons) have 1 representative mention.

- semantic_class and genericity are assigned to all representative mentions (and not to any other mention)?

- only the first occurrence of deictic expressions are marked as exophoric, all the other mentions referring to the same entity are marked as anaphoric.

7 Version Tracking

We keep track of the revision details of this guidelines in this section.

Version	Description
v1.0	The first version of the guidelines. This version describes the instructions for annotating the complete reference chains containing at least one third person singular pronoun.
v2.0	This version describes the instructions for annotating all referential expressions and reference chains in the text.
v2.1	The decisions made during the annotation process of complete chains are added to the guideline.

Table 1: Version details