Parallel nominal coreference annotation guidelines

Version 1.0

Yulia Grishina and Manfred Stede

Applied Computational Linguistics
FSP Cognitive Science
University of Potsdam

grishina|stede@uni-potsdam.de

---

1 Introduction

These guidelines present instructions for the annotation of nominal coreference in multilingual texts. They are based on the adaptation and extension of the German part of the Potsdam Coreference Scheme (PoCoS) (Krasavina and Chiarcos, 2007), but deviate from it in a few points. Moreover, they take into account the annotation conventions of the English part of the OntoNotes coreference scheme (Hovy et al., 2006). For the time being, these guidelines address only nominal *identity* coreference; event anaphors or abstract anaphors are not being annotated. Likewise, pleonastic pronouns and pronouns with no specific antecedent are excluded from the annotation. For the annotation, we use freely available MMAX-2 annotation tool[1].

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 selected for 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.

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] http://mmax2.sourceforge.net

1. FULL NOMINAL PHRASES, e.g. *a big blue sky;*

2. PROPER NAMES AND TITLES, e.g. *Mr. Black;*

3. PRONOUNS:

   - PERSONAL PRONOUNS (FIRST, SECOND AND THIRD-PERSON):

     We only annotate personal pronouns if they have a specific referent in the text.

     (1) a. Hello, can [I]$_1$ help [you]$_2$? - [Daisy]$_1$ asked [the lady]$_2$.
     b. Hallo, kann [ich]$_1$ Ihnen helfen? - fragte [Daisy]$_1$ die Dame.

     (2) a. If *you* need more information about *your* medical condition, read the Package Leaflet.
     b. Wenn *Sie* weitere Informationen über *Ihre* Krankheit oder deren Behandlung benötigen , lesen *Sie* bitte die Packungsbeilage.

     In example (1), we annotate the first-person pronoun *[I]* as referring to the specific antecedent *[Daisy]* and the second-person pronoun *[you]* as referring to the specific antecedent *[lady]*. In example 2, we do not annotate the personal pronouns *you* and *your*, because they do not have any specific antecedent in the text but refer to the abstract reader.

     We also do not annotate first-person pronouns if they denote the author of the text (and do not have a specific antecedent in the text):

     (3) *I* am sure, *our* time for standing pat, for protecting narrow interests and putting off unpleasant decisions - that time has surely passed.

   - DEMONSTRATIVE PRONOUNS

     (4) a. You need [a camera]$_1$ [that]$_1$ works in the dark. Hm, take [this]$_1$.
     b. Sie brauchen [ein Modell]$_1$, [das]$_1$ auch nachts funktioniert. Hm, nehmen Sie [dieses]$_1$.

     In example (4), the demonstrative pronoun *[this]* points back to it antecedent *[a camera]* mentioned in the previous sentence and must be annotated. Keep in mind that we do NOT annotate event coreference, that is why we do not consider demonstrative pronouns if they refer to a verb phrase or to a bigger discourse unit, as in the following example:

     (5) The London G-20 meeting recognized that *the world's poorest countries and people should not be penalized by a crisis for which they are not responsible.* With *this* in mind, the G-20 leaders set out an ambitious agenda for an inclusive and wide-ranging response.

     In example (5), *this* does not have a specific referent, but refers to the whole subordinate clause of the previous sentence and therefore should not be marked.

     Predicative constructions are annotated in the following way:

     (6) a. [This]$_1$ is a bank, but [it]$_1$ is not very well-known.

- RELATIVE PRONOUNS, such as *who*, *whom*, *whose*, *which*, *that* etc.

  (7) a. [The Army]$_1$, [which]$_1$ recruits heavily in the Punjab, will not use [their]$_1$ force there in the way [it]$_1$ is doing in the tribal areas.
  b. [Die Armee]$_1$, [die]$_1$ einen Großteil [ihrer]$_1$ Soldaten im Punjab rekrutiert, wird dort nicht mit Gewalt vorgehen, so wie [sie]$_1$ es in den Stammesgebieten tut.

  Keep in mind that pronouns can be ambiguous:

  (8) For both India and Pakistan, Afghanistan risks turning into a new disputed territory, like [Kashmir]$_1$, [where]$_1$ the conflict has damaged both countries for more than 50 years.

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

  In example (8), *where* is a relative pronoun and refers to *Kashmir* (to show this, one can substitute *where* by *in which*). Conversely, in (9), *where* is not a relative pronoun and should not be annotated.

- REFLEXIVE PRONOUNS

  (10) a. It's beginning to rain! - [Daisy]$_1$ exclaimed to [herself]$_1$.
  b. Es fängt an zu regnen! - sagte [Daisy]$_1$ zu [sich]$_1$ [selbst]$_1$.

  For German, reflexive pronouns must be annotated only if they are independent constituents, but not part of a reflexive verb:

  (11) Ich habe *mich* gestern gewundert. (*Mich habe ich gestern gewundert)

  (12) Ich habe [mich]$_1$ gestern gesehen. (Mich habe ich gestern gesehen)

  The following test should be applied: if the position of the reflexive pronoun can be changed, then the pronoun is an independent unit (12), otherwise it belongs to the verb (11). Reflexivity in German and in Russian can also be marked by other units, such as *selbst*, *selber*, *persönlich* that also must be annotated.

- PRONOMINAL ADVERBS (GERMAN)

  (13) Viele Amerkaner haben Probleme mit [Rassismus]$_1$; doch wir sind [dagegen]$_1$ immun.

  (14) The Army, which recruits heavily in [the Punjab]$_1$, will not use force [there]$_1$.

- HIS/HERS, HIS OR HERS 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.

5. NOMINAL PREMODIFIERS

   In case of English nominal premodifiers, we only annotate a nominal premodifier if it can refer to a named entity (*[the [US]*$_1$ *politicians]*$_2$) or is an independent noun in the genitive form (*[[creditor's]*$_1$ *choice]*$_2$); in all other cases, nominal premodifiers are not annotated as separate markables (*bank account*).

6. GENERIC REFERENCE

   Generic nouns can co-refer with definite full NPs or pronouns, but not with other generic nouns. For example:

   (15) a. [Computers]$_1$ are expensive. But [they]$_1$ are really useful. *Computers* cost a lot of money.
   b. [Computer]$_1$ sind teuer. Aber [sie]$_1$ sind richtig nützlich. *Computer* kosten viel Geld.

   In this case, we only link the anaphoric pronoun *[they]* to its antecedent in the first sentence, *[computers]*, but we do not annotate the generic noun *computers* in the third sentence.

7. GROUPS

   If all elements from a group are referred to by an anaphoric pronoun, create a group markable consisting of the set elements and then link the anaphoric pronoun to it.

   (16) Did [your husband]$_1$ buy Lorna, [Mrs. Humphries]$_2$? - No, [we]$_{1+2}$ bought her together.

8. TEMPORAL EXPRESSIONS

   Temporal expressions are to be annotated if they co-refer.

We do **NOT** annotate predicative forms: When a copula is used to "equate" two nominal expressions, the predicated one is not a markable:

(17) a. [Oxford]$_1$ is *a university.* [It]$_1$ has a long history.
   b. [Oxford]$_1$ ist *eine Universität.* [Sie]$_1$ hat eine lange Geschichte.

Keep in mind that in the case of the change of perspective on the referent of an anaphoric expression we should start a new chain if the already mentioned referent becomes unspecific. See the following example:

(18) So [Daisy]$_1$ tried to turn it off but pushed the wrong button and the whirring sound increased. At this point Pam's ex-husband became aware of it and turned round furiously. He realized [someone]$_2$ was watching him and swore profusely. Then he made towards [Daisy]$_2$ as though to hit her.

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;
- deverbal modifiers (participial constructions, regardless whether in pre- or postposition) that can be substituted by a subordinate clause, for example:

  (19) [Regional conflict, involving all of the region's states and increasing numbers of non-state actors]$_1$, has produced large numbers of [trained fighters, waiting for the call to glory]$_2$.

  In this case, both *[regional conflict, involving all of the region's states and increasing numbers of non-state actors]* and *[trained fighters, waiting for the call to glory]* are markables.

- dependent prepositional phrases (for example, *[Queen of England]*$_1$).
- appositions, i.e., additive material that is not syntactically integrated, are included into the markable span, but are not annotated separately:

  (20) a. [JuD, Party of Proselytizing,]$_1$ was founded in 1972.
  b. [Jud, Partei der Missionierung,]$_1$ wurde 1972 gegründet.

However, full clauses, in particular relative clauses, are not taken as parts of the markable rooted in the NP head. Therefore we annotate relative pronouns separately (see 2.1).

3 Annotation process

The annotation process selects only those nominal expressions ('markables') that actually appear in a coreference chain (i.e., those that are mentioned at least two times in the text). When some entity is mentioned only once by some referring expression (a so-called 'singleton'), this expression is *not* a markable. Therefore, the annotation process involves a certain amount of "going back and forth" in the text. Moving from left to right, when you encounter a referring expression R, check whether it anaphorically refers to an entity that has already been mentioned. If that is the case, establish a markable for R and link it to its nearest antecedent, i.e. the most recent expression A that has the same referent. If A is already a markable – that is, it already participates in a coreference chain –, this can be done right away. If, on the other hand, A is the first mention of that referent, then A also has to be established as a markable before the coreference link can be established. In the 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').

4 Attributes

4.1 Attributes for all markables

4.1.1 referentiality

1. not_specified - during the annotation, no decision was taken
2. discourse_cataphor - newly mentioned underspecified discourse entity which meaning is denoted later by a more specific antecedent (except for bridging-contained)
3. referring - discourse entity that can interpreted on the basis of the previous context
4. discourse_new - the first mention of a discourse entity + all bridging markables, including bridging-contained, except for cataphoric (to be marked as such)
5. other - impossible to decide between (a)-(d).

4.1.2 dir_speech

1. text_level - the markable does not appear in direct or indirect speech
2. direct_speech - the markable appears in the direct speech
3. indirect_speech - the markable appears in the indirect speech.

4.1.3 phrase_type

1. np - nominal phrase (in general, markables are nominal phrases)
2. pp - prepositional noun phrase (used only in case of contractions, when we have to annotate the preposition as a part of the markable)
3. other - if (a) and (b) are not applicable.

4.1.4 np_form

1. NE - named entity
2. defNP - definite NP
3. indefNP - indefinite NP
4. ppers - personal pronoun
5. ppos - possessive pronoun
6. padv - pronominal adverb
7. pds - demonstrative pronoun
8. rel - relative pronoun
9. refl - reflexive pronoun

4.1.5 ambiguity

This attribute is not applicable.

1. not_ambig
2. ambig_ante
3. ambig_rel
4. ambig_rel$_a$nte
5. ambig_idiom
6. ambig_expl
7. ambig_other

4.1.6 complex_np

Complex NPs are those containing embedded NPs, deverbal modifiers or appositions (as defined in 2.2).

1. not_specified - during the annotation, no decision was taken
2. yes
3. no

4.1.7 grammatical role

1. not_specified - during the annotation, no decision was taken
2. sbj - subject
3. dir_obj - direct object
4. indir_obj - indirect object
5. other - none of the above applicable.

4.1.8 comment

The comment field is to be filled in case of any uncertainty regarding the aforementioned attributes.

References

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). OntoNotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60. Association for Computational Linguistics.

Krasavina, O. and Chiarcos, C. (2007). Pocos: Potsdam coreference scheme. In *Proceedings of the Linguistic Annotation Workshop*, pages 156–163. Association for Computational Linguistics.