Clinical Coreference Annotation Guidelines (with excerpts from ODIE guidelines and modified for SHARP)

Arrick Lanfranchi and Kevin Crooks

The following is a proposal/summary of the ODIE guidelines with some references to the OntoNotes coreference annotation guidelines. Completed guidelines will be provided after pilot annotation and further discussion. Our goals are:

- to make the coreference annotation task and guidelines as transparent as possible for non-domain annotators
- to narrow the scope of the ODIE guidelines to coreference annotation only, with automatic extraction from other annotation layers of extra-informative features included in the ODIE coreference annotation task.

The University of Colorado coreference annotation team is proposing to adapt both the OntoNotes v. 7.0 (2007) and ODIE (2010) guidelines as needed to accommodate coreference annotation in the clinical domain.

What is coreference annotation?

The purpose of coreference annotation is to link all the specific mentions in a text that refer to the same entities and events, and to distinguish between types of co-reference. Entities eligible for co-reference includes nouns, noun phrases, pronouns, and nominalized verbs, and the relations identified are identity chains and apposition. (See OntoNotes guidelines, revised 10/11/07, pg. 2)

MARKABLES:

All nouns, noun phrases (including relative clauses), nominal modifiers, pronouns, and nominalized verbs are considered markables eligible for coreference relation annotation. The data we will be using will come pre-annotated with markables. We are not restricted to these markables only. We can create our own markables. Note, however, that if the span of a pre-annotated markable is not complete, rather than creating a new markable, we should edit the span of the pre-annotations (this will occur frequently with articles “the” and “a”, and the possessive ‘s). When expanding a markable select the span that contains the head and as much of the modifying information as possible.

Only markables that participate in chains or in appositives are to be annotated. All appositives will be annotated regardless of whether or not they participate in an identity chain. Annotate the longest and the most specific span that you think belongs to the markable linguistic expression, which includes determiners as well as modifying information to the head noun such as prepositions and relative clauses.

Example (ODIE data set m1, document 112831390_2):

(1) “HHHHHHH was seen by [Dr. DDDDDD in neurology]… I have spoken to [Dr. DDDDDD who is now at SSSS Clinic-SSSSSSSSSS]”

“Dr. DDDDDD in neurology” and “Dr. DDDDDD who is now at SSSS Clinic-SSSSSSSSS” are taken as the “longest and the most specific span(s),” as they are treated as a noun phrase constituents according to Penn Treebank policy.

Nominalized verbs can be annotated as markables if they participate in an identity chain. In these cases, the markable should include only the single-word verb.

Example:
(2) The patient will need a CBC as long as he is on azathioprine to (M1 monitor) for leukopenia. The patient wants to have his physician arrange for (M2 this monitoring).
(3) She is encouraged to (M1 walk or climb) stairs but should avoid physical activity more extensive than (M2 this).
(4) The glucose does not directly affect potassium levels but should be given to (M1 prevent) hypoglycemia. (M2 This) is also a temporary measure.

In all 3 examples, M1 and M2 are coreferrential.

Try to only annotate verbs if replacing the coreferring markable (usually a demonstrative, *this* or *that*) with the nominalized form of the verb does not change the meaning, and have a clear co-reference. In this case, it is obvious that "this" corefers to both walking and climbing, and we can replace it with "This walking or climbing" without losing the coreference, and thus we select both verbs in the span of annotation.

Dates and locations are considered to be atomic, and nested dates and locations cannot be extracted. For example, in "November 7, 2000", there can be no co-ref chains for "November" or "2000". The same applies for locations such as "Boulder, Colorado, USA".

Example:
(5) (M1 11 May 2006) 10:11AM Exam: CTA Lower Ext. ORIGINAL REPORT: (M2 11 May 2006).

M2 and M3 are coreferrential.

Dosages are also considered to be atomic.

Example:
(6) She was started on a course of (M1 prednisone 40-mg).

The entire phrase "prednisone 40-mg" would be one markable, and is eligible for coreference if there are other mentions of that specific medication.

Example:
(7) Patient receives (M1 (M3 Lasix 40-mg p.o. q.d.) and (M4 propranolol 50-mg p.o. q.d.)) We will continue him on (M2 these doses of antihypertensives), provided he tolerates (M5 the beta blocker) as well as (M6 the diuretic).

M1 corefers to M2, M3 corefers to M6, and M4 corefers to M5. However, "Lasix" and "propranolol" cannnot be extracted as markables and thus will not corefer.

Since we are treating stative adjectives as predicates, occasionally, you will encounter one that corefers.

Example:
(8) Heart rate was (M1 normal). (M2 This) is acceptable.

Nested NPs and Premodifiers:
Overlapping annotations of markables are allowed if they are a part of a chain. Adjectives, determiners and other modifiers are to be included in the span if relevant.

Example (ODIE Coreference guidelines, Jan. 5, 2009):
(9) The patient was transferred to IIIIIIIIII for (M1 explantation of (M2 (M6 a pacemaker) system)). The patient underwent (M3 the procedure) without any complications. On **DATE[], (M4 the pacemaker) was (M7 explanted) from the left shoulder.

In the above example, M1 and M6 have overlapping text spans, however each participate in identical chains – M1 and M3; M4 and M6. Both M1 and M3 (along with M7) are to be annotated as well as M4 and M6.

Premodifiers may be annotated as a markable without their head nouns, but only if they participate in identity chains. For example:

(10) I think that the patient is metabolically indeterminate in terms of (M1 stone) formation and growth. Magnesium is elevated which is protective against (M2 kidney stones).

M1 and M2 corefer, as both are general mentions of kidney stones. However, they cannot be linked to the mention of "stone" in the following passage from the same article:

(11) I compared it to a stone protocol CT-scan of Btac, 3881.

This is because we can only extract one level; that is, only the largest premodifiying span is extracted. In this above example, "stone" modifies "protocol", which in turn modifies "CT-scan", and thus we can only annotate "stone protocol" as a markable. We also would only annotate "stone protocol" as a markable if there are other individual, non-premodifying mentions of "stone protocol" in the same note.

Names are markables.

Example:
(12) I discussed my clinical expression at length with (M1 Mr. SSSSS) and (M2 his) wife. I have recommended (M3 he) apply DesOwen lotion b.i.d. prn.

"Mr. SSSSS" (M1), "his" (M2) and "he" (M3) are coreferrential.

Possessives are included in markables.

Example:
(13) (M1 MM. EEEEE'x) diet for many years was relatively low in calcium. (M2 He) tended to go easy on meat and moderate if not easy on salt.

M1 and M2 are coreferrential – we extract the possessive 's with the markable. In our data, personal information has been de-identified, but in this case it is clear that EEEEE'x is a possessive, and thus we include the 'x in the span.

See section 4.3., MUC-7 guidelines for more examples.

Occasionally there are mentions that are adjectival in nature but clearly corefer with a nominal mention within the text. *Pre-* and *Post-* terms that corefer are also annotated. These *pre-* and *post-* terms are eligible as members in an Identical chain only if there is a coreferring noun with which to form a chain. There may be other such prefixes that behave similary so this is not limited to *pre-* and *post-* terms. These occur in constructions like “preoperative”, “post-surgical”, and “post-partum”.

Example:

(14) Thus, Patient is ok for (M1 anesthesia) with the proviso that he be seen by cardiology pre- and postoperatively....Patient seen at the request of DD. DDDDD for (M2 pre-anesthetic) medical evaluation...

(15) Stable (M1 postoperative) course. She has only noted minimal change in her hand function following (M2 the procedure).

M1 and M2 are coreferring in both examples.

De-identified mentions are to be annotated as markables and subsequently as members of a coreference pair/chain if it is clear that they are truly coreferring. There are multiple schema used for de-identification, but all should be treated in the same way when it is obvious they co-refer.

Example:

(16) She certainly could try aspirin which in a study conducted by (M1 Sx. Yxgwx Gqxlmwzwi) seemed to be effective at least partially. (M2 Sx. Yxgwx Gqxlmwzwi) has presented this data.

M1 and M2 are coreferring.

Interesting cases:

- Disjoint spans

Example:

(17) Two dimensional echocardiology: (M1 Segmental left ventricular function). Final Impression: (M2 Normal left ventricular) size and (M2 function).

In this example, we would want to corefer “segmental left ventricular function” with “normal left ventricular… function” (M1 corefers with M2). The latter is a disjoint span. Knowtator allows for the disjoint span annotations.

- Age

Example:

(18) Mr. Smith, 60, presents today…

We would here create an appositive with “Mr. Smith” as the head, and “60” as the attribute. Although these two markables are not technically equatives, this is considered to be an elliptical construction of “Mr. Smith, who is age 60”, and we need to capture that information.

- Post-modifiers

Example:
(19) (M1 (M2 Kidney, left), nephrectomy)

M1 illustrates the full span of "nephrectomy" which can be translated as "left kidney nephrectomy". As such, M2 is the full premodifying span which is linked to other coreferring equivalents of *left kindey*.

- Section numbering

In pathology notes, sub-sections are frequently headed by a numbering system. We will ignore these numbers.

Example:
(20) #1 Skin rash

"#1" is not eligible for coreference.

- Attributes and Syntactic Equatives[1]

In the case of Syntactic Equatives, also referred to as copular constructions, we only annotate the left-most markable (adopted from OntoNotes).

Examples:
(21) (M1 Mr. Callahan) is (M2 the president of IBM).
(22) (M1 Mr. Smith) is (M2 a 89-year gentleman).

Only M1 in each case is annotated as a member of the chain.

RELATIONS:
The RELATION TYPE attribute indicates the relation between annotated markables. MUC-7 task annotates for the IDENTITY relations only. The relations we are annotating for are:

- Identity (or coreference)
- Appositive
- Set/subset
- Part/whole

Set/subset and part/whole are currently out of scope for this project, but may be added soon.

1[1] The SHARP team is performing several layers of annotation (POS tagging/Treebanking, UMLS, Propbank) as well as coreference annotation. The output of these layers can be used to automatically extract most Attributes when not already inherent in the data (such as definite and indefinite noun phrases). SHARP proposes that the coreference task focus only on the annotation of coreferential relationships with the Attributes being extracted automatically. SYNTACTIC EQUATIVES: This category is similar to appositives except that the two coreferring items are separated by an equative. In addition to NER, the amount of processing required for this class is the ability to identify equatives like "of", "is", etc. This information is easily extractable based on Propbank frames for "be" and/or tree structure (for small clauses and secondary predicates) and easily distinguishable from appositives.

Two markables have an **IDENTITY** relation, or corefer, if they refer to one and the same (discourse) referent. Following the MUC-7 specifications, the IDENTITY relation has several important semantic characteristics. The Identity relation is symmetrical (if A is IDENT to B, then B is IDENT to A). It is also transitive (if A is IDENT to B and B is IDENT to C, then A is IDENT to C, and C is IDENT to A). The IDENT relationship is not directional to set it apart from part-whole and set-subset relations.

Example:
(23) (M1 Mr. Smith) complained of a headache. (M2 He) also had a sore throat.
(24) Mr. Smith (M3 ran). I saw (M4 it).

The relation between M1 and M2 is Identity. The relation between M3 and M4 is Identity.

Example:
(25) (M1 Aortic root): (M2 2.9 cm) (2.0-3.7cm) (M3 The aortic root size) is normal.

M2 and M3 is the only coreference pair in this example.

Two markables have an **APPOSITIVE** relation if two NPs having the same semantic meaning occur adjacent to one another, separated only by punctuation – almost always a comma, colon, dash, or parenthesis. Appositives are treated as a separate relationship, where the most specific markable, determined with a specificity hierarchy, is the head and the less specific is the attribute. Appositives can then be included in a chain, however, all appositives are annotated regardless of whether they corefer.

Examples:
(26) (M1 ZZZZZZ, MD, FFF), (M2 Pathologist).

M1 would be the head, M2 would be the attribute, which constitute one Appositive relationship. That Appositive relationship is then added to the identical chain for "ZZZZZZ, MD, FFF".

(27) (M1 37.00C) ~~(~~(M2 98.60F)~~)~~

M1 is the head, M2 is the attribute. Measurements are frequent in this data, and unit conversions like this are considered appositives.

(28) (M1 Consulting Surgeon): (M2 CCCCCCC, CCCCCCC H. Z.Q.)

M2 is the head, and M1 is the attribute, following the **SPECIFICITY HEIRARCHY:**

(See OntoNotes guidelines, revised 10-11-07, sect. 3.1)
Proper noun > pronoun > definite NP > indefinite specific NP > non-specific NP

If both nominals have the same level of specificity, select the left-most as the head. In some cases, appositives may include multiple markables. Select the most specific as the head, and the other markables as the attributes (See OntoNotes guidelines, revised 10-11-07, sect. 3.1).

Examples:
(29) (M1 AUTHOR) (M2 Contributing Author) (M3 AAAAAAAAAAAAAA. M.D.)

AAAAAAAAAAAAAA. M.D. would be the head, and both “AUTHOR” and “Contributing Author” would be the attributes.