Anaphoricity Annotation Guidelines for the Clinical Domain

Authors:
Guergana Savova, PhD
Wendy Chapman, PhD
Jiaping Zheng, M.S.

In consultation with
Lynette Hirschman, MITRE
Cheryl Clark, MITRE
Kevin Cohen, University of Colorado
Arrick Lanfranchi, University of Colorado
Martha Palmer, University of Colorado

The work was funded by grant R01 CA127979. We are thankful to our annotators, Donna Ihrke, Pauline Funk and Melissa Castine. The work was conducted under IRB 08-007020 and REN09050055/PRO07070252.

Table of Contents

1 General information and background 3
2 Scope of anaphoric relations 3
3 Annotation tool 4
4 Some definitions 4
5 Markables 4
5.1 Attributes 6
5.1.1 Phrasal tags 6
5.1.2 Type 7
5.1.3 Function tag 14
6 Pairs 16
6.1 Attributes 16
6.1.1 Anaphor 16
6.1.2 Antecedent 16
6.1.3 Bagga class 18
6.1.4 Pair relation type 21
7 Chains 24
7.1 Attributes 25
7.1.1 Pairs 25
7.1.2 Chain Relation Type 25
References: 25

1 General information and background

The guidelines described in this document closely follow the MUC-7 Coreference Task Definition (MUC-7, 1997). The modifications reflect the necessary adaptations to the clinical domain and its language especially as related to the type of the markable linguistic expression.

Our goal was to devise an annotation schema which can accommodate the annotations of a corpus which can subsequently be used for training and testing a number of anaphora resolution algorithms and models, e.g. mention-pair and entity mention models (Yang et al., 2008). *Anaphoric relations* (a.k.a. anaphoricity) are such relations between linguistic expressions where the interpretation of one of the linguistic expression (the anaphor) relies on the interpretation of another linguistic expression (the antecedent). Anaphoric relations can define identity, set/subset, part/whole relations between the participating linguistic expressions. *Coreferential relations* (or coreference) are anaphoric relations of type identity. Two expressions are coreferential if they refer to one and the same discourse referent. The form of the linguistic expressions includes noun phrases, verbs, adverbials, clauses. MUC-7 task linked coreferring relations between nouns; verbs and clauses were out-of-scope.

Identifying anaphoric textual mentions allows composite information extraction and structured template filling. For example, if a patient is diagnosed with colon cancer, which is referred to as “colon cancer”, “tumor”, “the disease”, the attributes pertaining to all these mentions are to be added to one data structure that describes the patient’s disease providing a unified codification to a domain-specific ontology.

The material for our task is clinical free text comprising of radiology, pathology and clinical notes across two institutions (University of Pittsburgh Medical Center and the Mayo Clinic). Clinical free text has characteristics that set it apart from other types of biomedical text requiring unique modeling of the domain.

2 Scope of anaphoric relations

Our task is to annotate a select set of anaphoric relations within a given document and across the paragraphs and the sections of that document. Cross-document coreference is out-of-scope.

Section headers are to be annotated if the information in them is relevant to anaphoric relations. An example of a header NOT to be annotated is the following:

S_O_H
Counters Report Type Record Type Subgroup Classifier
1,01TdvtyYejbW DS DS 1504
E_O_H
[Report de-identified (Safe-harbor compliant) by De-ID v.6.14.02]

An example of a header to be annotated is below where "**NAME[AAA, BBB M]" is annotated and later coreferenced with mentions of "the patient":

**INSTITUTION
CARDIOLOGY
DISCHARGE SUMMARY
PATIENT NAME: **NAME[AAA, BBB M]
ACCOUNT #: **ID-NUM
**ROOM
ATTENDING PHYSICIAN: **NAME[YYY M ZZZ]
ADMISSION DATE: **DATE[Sep 27 2007]
DISCHARGE DATE: **DATE[Oct 07 2007]

3 Annotation tool

The tool that we are using for this task is the Knowtator[1] as it is embedded in Protégé which allows a quick and easy way to build an annotation schema. The Knowtator assigns unique IDs to each created annotation which can later be used for unique instance identification.

4 Some definitions

See section 4.2 from MUC-7 guidelines for the definition for extensional descriptor, intensional descriptor and the grounding instance of a coreference chain.

5 Markables

Markables are the linguistic expressions to be annotated for this task. The MUC-7 task annotated the relations between elements of the following categories: NOUNS, NOUN PHRASES, and PRONOUNS. Note, that just because an element is a markable, it does not mean that there are later references to it as the markable may or may not participate in anaphoric relations. MUC-7 considered interrogative "wh-" NPs as non-markable, for example "Who is your boss?".

As MUC-7 guidelines state "the relation is marked only between pairs of elements both of which are markables. This means that some markables that look like anaphoric will not be coded, including pronouns, demonstratives, and definite NPs whose antecedent is a clause rather than a markable."

We extend the MUC-7 guidelines to add CLAUSE to the markable categories in order to capture coreferential relations between events.

Example:
(M1 Peter ran). I saw (M2 it).

[1] http://knowtator.sourceforge.net/

This would be an S in treebanked data.

M2 refers to M1 as what is stated in the sentence is that the speaker saw an event described in the first sentence.

Only markables that participate in anaphoric pairs and chains are to be annotated. Annotate the longest and the most specific span that you think belongs to the markable linguistic expression.

Example:
The patient was transferred to **INSTITUTION for (M1 explantation of a pacemaker system).

The text span for M1 associated with the most specific concept is to be annotated ("explantation of a pacemaker system") rather than the shorter span that relates to a more general concept ("explantation").

Overlapping annotations of markables are allowed if they are a part of an anaphoric pair or chain.

Example:
The patient was transferred to **INSTITUTION for (M1 explantation of (M2 a pacemaker system)). The patient underwent (M3 the procedure) without any complications. On **DATE[Oct 4 2007], (M4 the pacemaker) was explanted from the left shoulder.

In the above example, M1 and M2 have overlapping text spans, however both participate in anaphoric pairs and chains – M1 and M3; M2 and M4. Both M1 and M2 are to be annotated.

Names are markables.

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

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

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

Date expressions are also treated as atomic although we are not going to annotate coreferring temporal expressions, they are out-of-scope for this project.

Deidentified mentions of INSTITUTION are not to be annotated as markables and subsequently members of a coreference pair/chain because the deidentification masks whether they are truly coreferring.

Interesting cases:

- Disjoint spans
  Example:

  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. The Knowtator allows for the disjoint span annotations.

Syntactic expletives, or the existential *it*, are non-referential pronouns.

Example:
It is raining.

"it" is a syntactic expletive and is not anaphoric as its meaning does not depend on the interpretation of another markable.

The span of the markable is determined by the textual string that best and most specifically represents the shared referent. Adjectives, determiners and other modifiers are to be included in the span if relevant.

Example:
(M1 TWO DIMENSIONAL ECHOCARDIOLOGY): (M2 This) was (M3 a technically difficult study.)

The textual span of M1 includes the modifiers of "echocardiology" – "two" and "dimensional". The textual span of M3 includes the determiner "a" along with the modifiers "technically" and "difficult".

In the spans "colon, rectum" and "rectum…. colon", "rectum" is annotated as a markable on its own.

5.1 Attributes

5.1.1 Phrasal tags

Phrasal tags are the grammatical categories of the spanned markable. The possible tags are:

- Indefinite noun phrases (NPindefinite)
- Definite noun phrases (NPdefinite)
- Bare noun phrases (BareNP), i.e. missing determiners (see example below)
- Demonstrative noun phrases (NPdemonstrative)
- Personal pronouns (ProunounPersonal)
- Possessive pronouns (PronounPossessive)

- Demonstrative pronouns (PronounDemonstrative)
- Relative pronouns (PronounRelative)
- Proper noun (NounProper)
- Clause

Example of a bare noun phrase:
Two dimensional echocardiology: (M1 Segmental left ventricular function).

The phrasal tag for M1 has a bareNP phrasal tag as it does not have any determiner preceding it.

Proper noun phrasal tag can be assigned markables of type People (see section 5.1.2 for Types). Names of diseases, drugs, procedures are not to be assigned the proper noun phrasal tag.

Non-existential "it" is a personal pronoun.

We do not mark empty traces and ellipses.

5.1.2 Type

Type is the named entity (NE) class, or semantic class, that the markable belongs to. The purpose of the Type attribute is to indicate the class to which the markable belongs to.

Only markables belonging to or coreferring with a markable belonging to one of the listed types below are to be annotated:

- People
- Disease/Syndrome
- Sign/Symptom
- Procedure, including Test procedure
- Anatomical Site
- Laboratory or Test Result
- Indicator, Reagent, or Diagnostic aid
- Organ or Tissue function

The definitions of each type is based on Bodenrieder and McCray, 2003 groupings of semantic types.

Definition of a disorder (definition for disease or syndrome from the UMLS): A condition which alters or interferes with a normal process, state, or activity of an organism. It is usually characterized by the abnormal functioning of one or more of the host's systems, parts, or organs. Included here is a complex of symptoms descriptive of a disorder. Any mention that belongs to this set of UMLS semantic types:

- Congenital Abnormality

- Acquired Abnormality
- Injury or Poisoning
- Pathologic Function
- Disease or Syndrome
- Mental or Behavioral Dysfunction
- Cell or Molecular Dysfunction
- Experimental Model of Disease
- Anatomical Abnormality
- Neoplastic Process
- Finding

Definition for Sign or Symptom (following the UMLS definition): An observable manifestation of a disease or condition based on clinical judgment, or a manifestation of a disease or condition which is experienced by the patient and reported as a subjective observation.

UMLS Semantic types:

- Sign or Symptom

Definition of procedure (following the UMLS definition): see below listed semantic types that belong to the Procedure semantic group

- Diagnostic procedure: procedure, method, or technique used to determine the nature or identity of a disease or disorder. This excludes procedures which are primarily carried out on specimens in a laboratory.

- Educational activity: An activity related to the organization and provision of education.

- Healthcare activity: An activity of or relating to the practice of medicine or involving the care of patients.

- Laboratory procedure: A procedure, method, or technique used to determine the composition, quantity, or concentration of a specimen, and which is carried out in a clinical laboratory. Included here are procedures which measure the times and rates of reactions.

- Therapeutic and Preventative procedure: A procedure, method, or technique designed to prevent a disease or a disorder, or to improve physical function, or used in the process of treating a disease or injury.

- Research activity: An activity carried out as part of research or experimentation.

- Molecular biology research technique: Any of the techniques used in the study of or the directed modification of the gene complement of a living organism.

Definition of lab or test result (following the UMLS definition): The outcome of a specific test to measure an attribute or to determine the presence, absence, or degree of a condition.

UMLS semantic types:
- Laboratory or Test results

Definition of indicator, reagent, or diagnostic aid (following the UMLS definition): A substance primarily of interest for its use in laboratory or diagnostic tests and procedures to detect, measure, examine, or analyze other chemicals, processes, or conditions.

Definition of anatomical site (following the UMLS definition): Includes the following semantic types:
- Anatomical Structure
- Body location or region
- body part, organ or organ component
- Body space or Junction
- Body substance
- Body system
- Cell
- Cell component
- Embryonic structure
- Fully formed anatomical structure
- Tissue

Definition of organ or tissue function (following the UMLS definition): A physiologic function of a particular organ, organ system, or tissue.

UMLS Knowledge Server Browsing procedure:
- Log in to this URL (ask Guergana for the username and password): https://kscas.nlm.nih.gov/cas/login?service=http://umlsks.nlm.nih.gov/uPortal/Login[2]
- After you are logged in, go to UMLS and Source View Tab and type the term that you want to search for (Figure 1). Prune by source to SNOMED Clinical Terms. Click ok for the search to start. The result of the search is in Figure 2.
- Click on the link for the term under the UMLS View (see Figure 3) and check whether the term falls into one of the allowed UMLS semantic types for our task. In the example below, "left ventricular function" is in SNOMED CT and has a UMLS semantic type of Organ or Tissue function (allowed), hence the text span is to be annotated as a markable.

[2] The UMLS Knowledge Server was retired January 30, 2011. Screenshots are from the retired server. The new server is https://uts.nlm.nih.gov/home.html

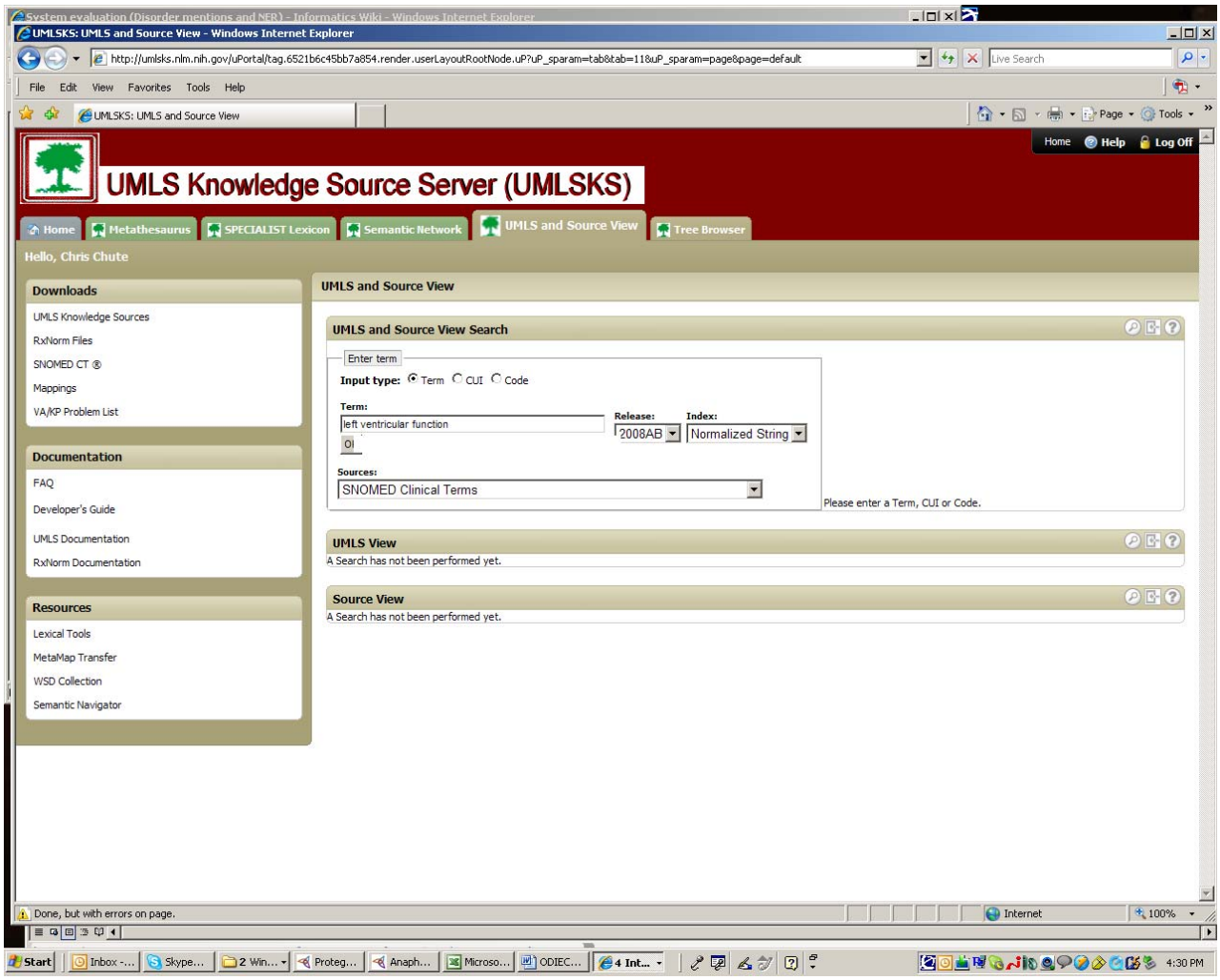

Figure 1: Screenshot 1

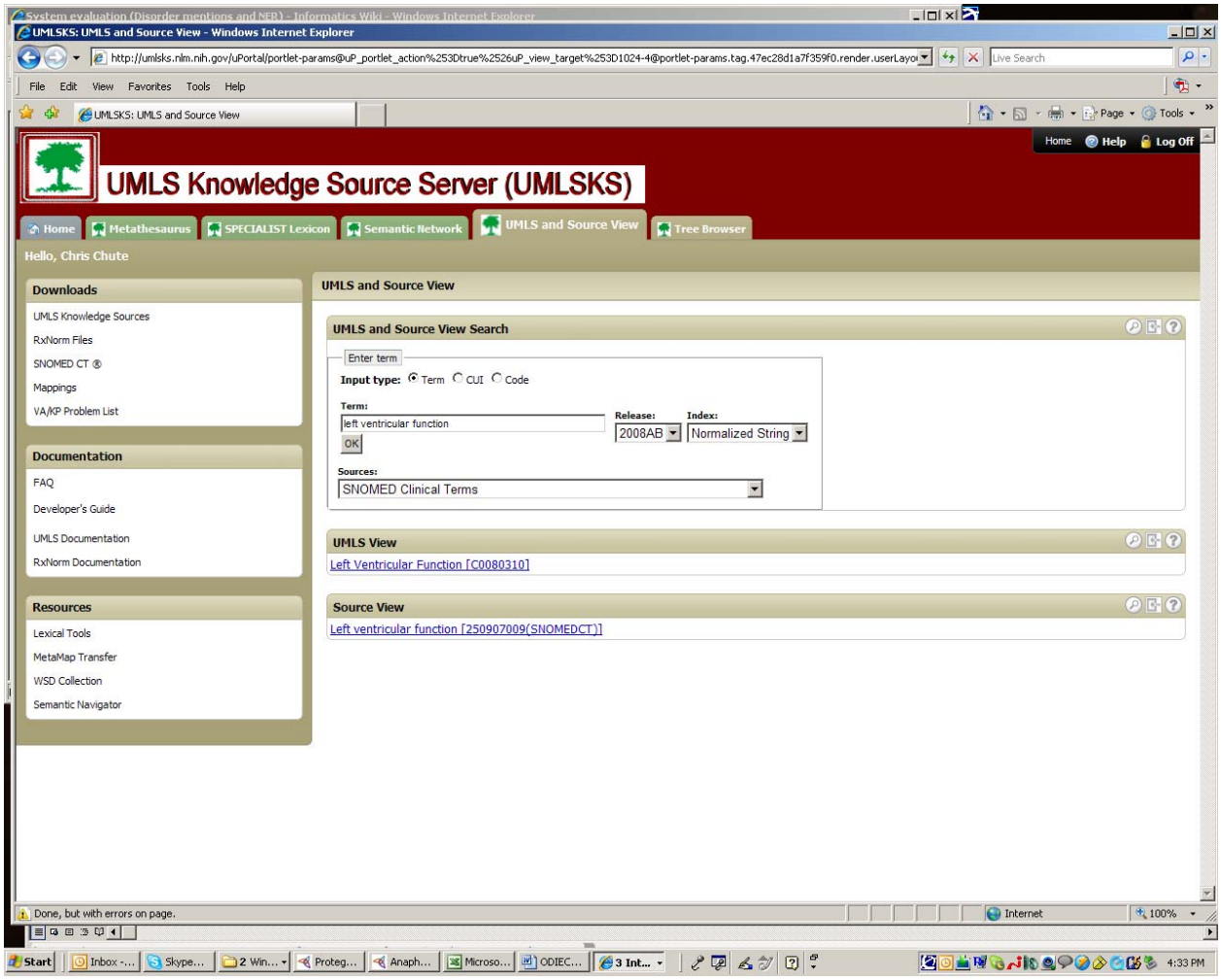

Figure 2: Screenshot 2

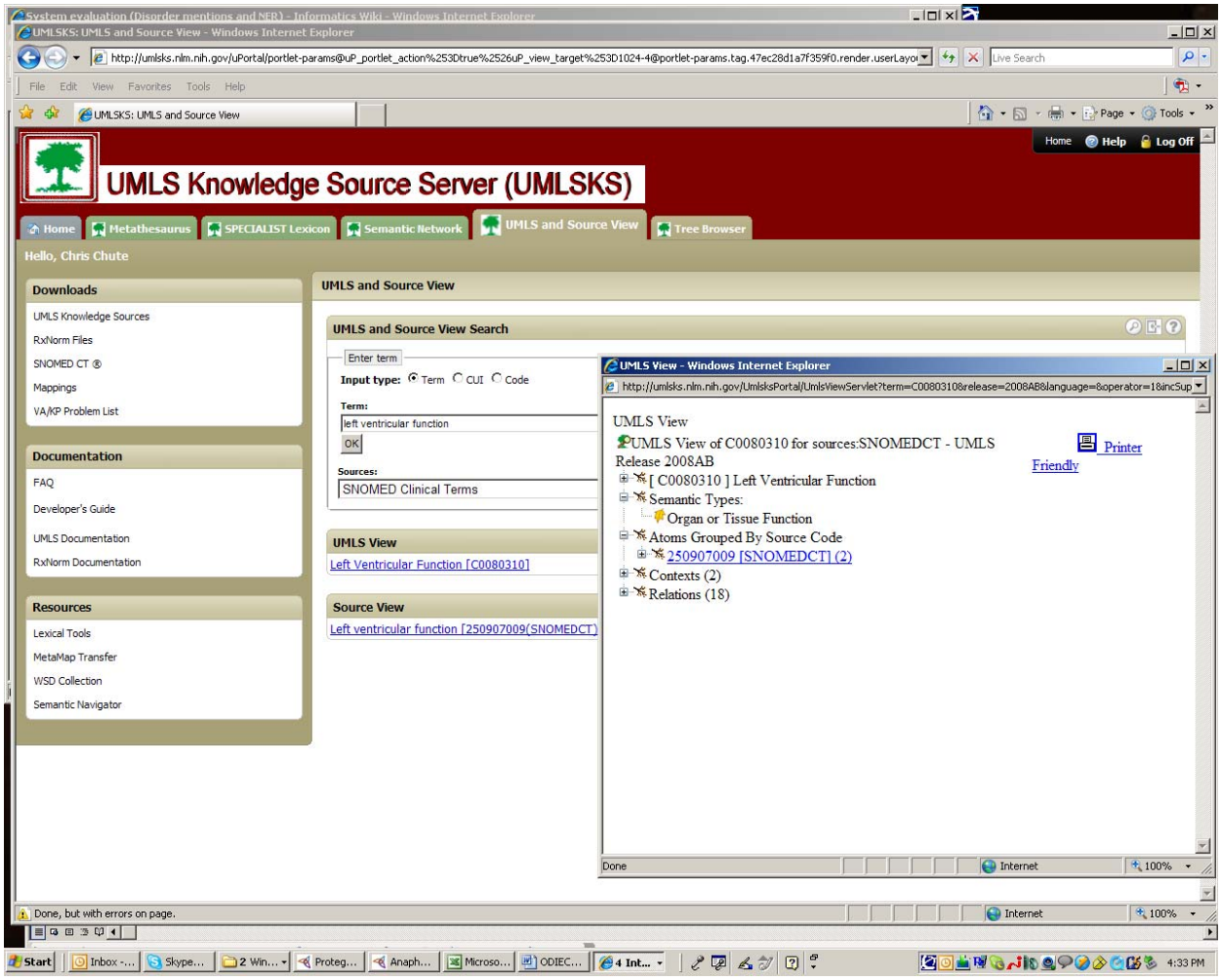

Figure 3: Screenshot 3

NOTE: **you are not to assign a specific CUI for the markable** which simplifies the task! As long as there is a match in SNOMED CT and it belongs to one of the allowed semantic types, then the text span is a markable.

Example:
"left ventricular size" has a UMLS type of Laboratory or Test Result.

If the semantics of the markable itself do not contain any type information, then assign the None type to that markable. For example, the pronoun "it" does not contain any particular type information by itself. If the semantics of the markable itself contain partial type information, then the context without the coreferring expression can be used to determine the final type assignment.

Example:
(M1 TWO DIMENSIONAL ECHOCARDIOLOGY): (M2 This) was (M3 a technically difficult study.)

M3 can be assigned the Procedure type as the section heading (TWO DIMENSIONAL ECHOCARDIOLOGY) is evidence for its type.

Of note, M2 is annotated as a markable because it refers to a markable (M1) that belongs to one of the allowable semantic types.

Example for the Other category is:
She was eating chicken two days ago when she felt (M1 a piece) stuck in her throat. She was unable to free the throat from (M2 this).

M1 is to be assigned the Other type as it clearly does not belong to one of the enumerated types. M2 is to be assigned the None type – the semantics of "this" does not have even partial evidence for its possible type. However, we are NOT annotating markables of type Other if they do not corefer with one of the allowed semantic classes.

Pronouns (except for personal pronouns) and demonstratives are to be assigned a None Type as they by themselves do not have a type. Rather, they inherit the type of their antecedents.

Example:
(M1 The cancer) grew rapidly. (M2 It) was surgically removed.

M2 is assigned a None type as it is a pronoun.

Example:
The patient was offered (M1 glucagon) (M2 which) she declined.

The type for M2 is None.

Example of types for disjoint spans:
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.

The phrasal tag for M1 and M2 is bareNP as neither has a preceding determiner. The NE type for M1 and M2 is Organ or Tissue Function (following the UMLS Semantic Types).

Proper nouns as part of disease names are not to be annotated by themselves but as a part of the disease mention.

Example:
The patient was diagnosed with (M1 Lou Gehrig's disease). (M2 The disease) progressed slowly.

M1 is a disease/disorder. “Lou Gehrig” by itself is not annotated as a Person.

Example:

(M1 FISH tests) show (M2 monosomy of chromosome 17 in 35% of the tumor cells). (M3 This FISH profile) is consistent with a diagnosis of a conventional clear cell renal cell carcinoma.

M1 is of type test procedure;
M2 is of type Lab or Test Result;
M3 is of type Lab or Test Result;

5.1.3 Function tag

Phrasal tags describe what the markable’s function in the sentence is.

- Surface subject (subjectSurface) follows the definition of PTB[3] annotations and marks the structural surface subject of both main and embedded clauses, including those with null subject. NPs which are the sole constituents in a sentence are marked as surface subjects, e.g. “1) Endoscopy.”, where the sentence consists of “Endoscopy” which is marked as a surface subject.

- Logical subject (subjectLogicalPassives) follows the definition of PTB annotations and is used to mark the logical subject in passives.

- Predicate Nominal subject (subjectPredicateNominal),
  Example:
  The patient is a 29-year old gentleman.

  where “a 29-year gentleman” functions as the Predicate Nominal subject

  Example:
  There is trace mitral regurgitation.
  where “trace mitral regurgitation” functions as the Predicate Nominal subject

- Indirect object (objectIndirect) – the affected participant in the event or the recipient of the direct object.

  Example:
  They sent him the test results.
  “him” is the Indirect object

- Direct object (objectDirect) – an object that is having something done to it.

  Example:
  They sent him the test results.

[3] http://bulba.sdsu.edu/jeanette/thesis/PennTags.html

"the test results" is the direct object.

- Prepositional object (objectPrepositional) – the object of a preposition

  Example:
  The medication was given to the patient.
  "the patient" is the Person markable and it is the object of the preposition *to*, and the indirect object of the verb *give*.

  Direct and Indirect objects take precedence over Prepositional objects if the markable is the object of a preposition.

  Example:
  (M1 They) sent (M2 the test results) to (M3 him).

  M3 is an indirect object and the object of a preposition. M3 gets the IndirectObject function tag.

- modifierToSubjectSurface, e.g.

  Example:
  The patient was diagnosed with (M1 colon cancer). The (M2 colon cancer) examination went fine.

  M2 gets the modifier to surface subject function tag

- ModifierToSubjectLogicalPassives

- modifierToSubjectNominalPredicate

- modifierToObjectDirect

- modifierToObjectIndirect

- modifierToObjectPrepositional
  Sometimes prepositional phrases are modifying objects of another prepositional phrase. In these cases, we want to capture the modifying relationship.

  Example:
  The patient was given a colonoscopy of (M1 the ascending colon).

  M1 is assigned the function tag of modifier to a prepositional object.

- Section Heading (sectionHeading) for these markables that are section headers,

  Example:

"Final diagnosis", "Procedure".

6 Pairs

Anaphoric relations are such relations between linguistic expressions where the interpretation of one of the linguistic expressions relies on the interpretation of another linguistic expression. *Coreference* is the identity relation between markables referring to the same entity. An anaphor is a linguistic expression which points to a previous item (Olsson, 2004), therefore a coreference relationship is a subset of anaphoric relationships. The Pair structure links the anaphor with its antecedent, identifies the class of relation between them (Bagga coreference classes) and the type of anaphoric relation between them.

If the coreference requires too much inferencing and domain knowledge, do not annotate the coreferring expressions. For example, the markables "overall left ventricular size" and "segmental left ventricular size" are likely to corefer, but arriving at that conclusion requires too much domain knowledge in addition to inferencing, hence these two markables are not to be annotated as coreferring.

6.1 Attributes

6.1.1 Anaphor

This slot is reserved for the anaphoric expression, i.e. the markable whose interpretation is dependent on the meaning of another preceding or subsequent markable. As Borthen, 2004 points out "anaphoric expressions often have an impoverished descriptive content that makes their meaning (more or less) underdetermined if they appear in isolation".

Example:
The patient was diagnosed with (M1 colon cancer). (M2 It) spread rapidly.

The instance "it" (M2) is the anaphor as it refers to a previous mention, in this case "colon cancer" (M1). M1 and M2 comprise an anaphor-antecedent pair.

6.1.2 Antecedent

Antecedent is the markable or linguistic expression that the anaphor refers to, or in other words, the markable on which the anaphor is dependent. In the previous example, M1 is the antecedent of M2. Options for selecting the antecedent:

Option 1: pick the immediately preceding coreferring markable (Borthen, 2004). This is the practice used in the computational literature (Gundel, personal communication).

Option 2: pick the first mention of a coreferring markable, or the grounding instance. This is the practice used in the linguistic literature (Gundel, personal communication)

For this project, our definition of antecedent is Option 1. As Borthen points out, the motivation for this definition of antecedents is that the closest antecedent candidate of an anaphor plays a particularly important role in the resolution process of anaphora. NOTE: Select the immediately preceding anaphoric markable to exclude possessive pronouns as antecedents if the pronouns are preceded by a coreferring proper noun in the same sentence (see the second example below)!!!!

Example:
The patient was diagnosed with (M1 colon cancer). (M2 The tumor) spread rapidly. (M3 It) was surgically removed. (M4 It) was 2 x 3 cm in size.

M1 is the antecedent of M2; M2 is the antecedent of M3; and M3 is the antecedent of M4.

Example:
(M1 Mr. Smith) is 89-year old gentleman who presents with a several year history of inverse psoriasis. (M2 I) discussed (M3 my) clinical expression at length with (M4 Mr. Smith) and (M5 his) wife. (M6 I) have recommended (M7 he) apply DesOwen lotion b.i.d. prn.

The antecedent of M7 is M4, not M5 as M5 is a coreferring possessive pronoun preceded by a proper noun in the same sentence.

Example:
Mr. Smith is a 57-year-old gentleman who presents with a several-year history of (M1 a recurrent, pruritic, burning eruption in the groin and navel). He states that he has tried multiple over-the-counter, anti-fungal preparations including Tinactin, Cruex, tolnaftate, and zinc oxide, all of which have been helpful to some degree. His family physician prescribed Nizoral cream which was also of some benefit. He states, however, that (M2 the eruption) often recurs after clearing. He has tried multiple approaches such as change in his diet and detergent thinking that this may have resulted in (M3 this eruption), however these have also been without benefit. He notes that when he was laid up in bed for 22000 weeks because of a back injury, (M4 the eruption) cleared. (M5 The eruption) also seems to improve when in the sun. He is referred to Dermatology for further evaluation and treatment.

Anaphor	antecedent
M2	M1
M3	M2
M4	M3
M5	M4

Example:

****

NAME: (**M1** **NAME[AAA, BBB])
MEDICAL RECORD #: **ID-NUM
SERVICE: **ID-NUM
SURGEON: **NAME[YYY ZZZ], MD
ASSISTANT(S): Virginia Kennihan, rn, **NAME[VVV UUU], RN

ASSISTING MD:
OPERATIVE DATE: **DATE[Dec 25 2007]
DICTATED **NAME[TTT]: **DATE[Dec 25 2007]
ADMISSION **NAME[TTT]: **DATE[Dec 25 2007]

TITLE OF OPERATION: Colonoscopy (CPT-45378)

ANESTHESIA: Midazolam 2 mg IV, Fentanyl 100 micrograms IV

PREOPERATIVE DIAGNOSIS(ES): (**M2** This **AGE[in 50s] year old female patient) is an acceptable candidate for colonoscopy. The indication for this procedure is screening for colorectal cancer and polyps.
****

The antecedent of the first mention of the patient is to be the name of the patient from the header if such is available. In this example, M2's antecedent is M1.

6.1.3 Bagga class

Bagga, 1998 proposes 11 coreference classes according to type of coreference and the amount of processing required to resolve it. The class is determined by the relationship between the anaphor and its antecedent. The anaphor is the anchoring point for making a decision about the Bagga coreference class. For example, if the anaphor is a pronoun, the Bagga class is "pronouns."

Example:
(M1 The patient) complained of a sore throat. (M2 She) has body aches as well.

The anaphor is "she" (M2) and the antecedent is "the patient" (M1). The Bagga coreference class is pronouns as the anaphor is a pronoun.

Here is a list of the classes:

- Appositives
  Where the anaphor is listed immediately after the antecedent, separated by a comma.

  Example:
  (M1 Louis Ferstner), (M2 President and CEO of IBM).

  The amount of processing required is the identification of important named entities in the text.

- 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.

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

The markables in these pairs are syntactic equatives.

- Proper names
  Example:
  (M1 President Clinton) gave an address yesterday. (M2 Mr. Clinton) praised the economy.

  In addition to NER, the amount of processing required for this class is the ability to generate syntactic variants of the NEs recognized.

- Pronouns
  All pronominal coreference placed in this category can be resolved using either linguistic principles or other syntactic methods.

  Example:
  (M1: Mr. Smith) is a 69-year-old-gentleman. (M2: He) complains of a sore throat.

  The Bagga type is “pronouns” as the anaphora is a pronoun.

- Quoted speech pronouns
  Pronouns used in quoted speech.

  Example:
  (M1: John) said: “(M2: I) am going to the mall.”

  In addition to the amount of processing required for the Pronouns class, this class also requires the ability to recognize quoted speech.

- Demonstratives
  This class includes coreference in which demonstrative phrases like “this”, “that”, etc. corefer with objects in the text.

  Example:
  He told me I had better handle this.

  The amount of processing required is similar to that of pronouns. However, this category requires more processing than pronouns because not all occurrences of “this” and “that” actually refer to another entity in the text (as in the example above where “this” could refer to a complex situation described in the discourse).

- Exact matches
  This does not include proper names.

Example:
two occurrences of the noun phrase “colon cancer”.

The amount of processing required for this class is noun phrase detection because coreferences requiring an exact string match almost always occur between two noun phrases.

- Substring matches
  This category consists of coreferences where all the words of one coreferring expression appear in another (the words need not be necessarily contiguous), such as “staph bacteremia” and “staph bacteremia infection”. As in the case of the Exact Match class, the amount of processing required is the ability to detect noun phrases.

- Identical lexical heads
  This category consists of coreferences where the two noun phrases share the same (morphologically identical) head – but the modifiers are different, such as “severe chest pain” and “pain”.

  Identical lexical heads can be considered a more specific case of substring matches. If both apply to a pair, pick the most specific class which would be the identical lexical heads.

  Example:
  “Thickened aortic valve” and “the aortic valve” are of Bagga class identical lexical heads with a lexical head of “valve”. Same applies to the pair “lung cancer” and “cancer in the lungs” where “cancer” is the lexical head for both markables. On the other hand, “staph bacteremia” and “staph bacterimia infection” are substring matches as they do not share a lexical head – the head of the former is “bacteremia” while the head of the latter is “infection”.

  In addition to noun phrase detection, the ability to recognize the head of each noun phrase is required for this class.

- Synonyms
  Coreferring expressions that do not meet any of the above definitions but have the same meaning.

  Example:
  Patient complains of (M1: shortness of breath). (M2: The dyspnea) has lasted for three days.

  A dictionary containing synonyms is required for this class to establish the synonymy relation between the textual strings of M1 and M2.

- External world knowledge

This category, in addition to including some very hard coreferences which required external world knowledge, also includes coreferences like “Hertz” and “the car rental company”.

Example:
(M1 John Doe)
SOCIAL HISTORY: (M2 The patient) drinks alcohol occasionally, is a nonsmoker.

In this example, M1 and M2 are coreferential, and external knowledge and/or some sort of reasoning is needed to identify the patient as John Doe.

- Ontology knowledge
  This class is in addition to the original Bagga classes and is to be subsumed by the External world knowledge class in the original Bagga classes. The Ontology knowledge class is added to determine the usability of an ontology for anaphora resolution in the clinical domain.

  Example:
  The patient is a 60-year-old gentleman who presented with complaints of shortness of breath and was found to have (M1 staph bacteremia). The patient was transferred to **INSTITUTION for explantation of a pacemaker system that was felt to be involved by (M2 infection).

  M1 and M2 are coreferential and the Bagga class is based on the ontological knowledge that staph bacteremia is a kind of infection.

6.1.4 Pair relation type

The PAIR RELATION TYPE attribute indicates the relation between the anaphor and its antecedent. MUC-7 task annotates for the IDENTITY relations only. The relations we are annotating for are:

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

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 IDEN 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:
(M1 Mr. Smith) complained of a headache. (M2 He) also had a sore throat. (M3 (M4 Mr. Smith) ran). I saw (M5 it).

The relation between M1 and M2 is Identity. The relation between M3 and M5 is Identity.

Example:

(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.

Example:

(M1 FISH tests) show (M2 monosomy of chromosome 17 in 35% of the tumor cells). (M3 This FISH profile) is consistent with a diagnosis of a conventional clear cell renal cell carcinoma.
(M4 Centromere Probe): (M5 CEP17) ; (M6 Copy 1: 21 (35%)); Copy 2: 39 (65%)

The coreference pairs in this example are: M2 and M3; M3 and M6.

In general, the above 2 examples fall in the category of function-value pairs. In this project, we will annotate coreferring relations between function-function and same function value-value (markables have to belong to the same category). We consider the function-value pair as a part of an information model. For example, consider the sentence "(M1 The temperature) was (M2 96F). (M3 It) dropped to (M4 79F).". There are 2 instantiations of the function-value pairs:

Temperature (time1, 96F)
Temperature (time2, 79F)

Mentions M1 and M3 refer to the general function information model, which triggers the instantiation of the two information models. Hence, we want to discover the coreferring relation between M1 and M3. The function/value discovery is then passed to the method for model population which is outside the scope of this project.

Are predicative NPs always coreferential with their subjects?
Example:
(M1 Mr. Smith) is (M2 a 65-old gentleman) who complains of a sore throat. (M3 Mr. Smith) is (M4 himself) today.

Borthen, 2004 argues that this "identity of reference" is not encoded through an anaphoric relation but through predication. The assumption is that "anaphoric relations are relations between constituents as such, where it is first of all features of these constituents (such as form, gender, lexical content, relative distance, etc.) that give signals about the intended anaphoric relation. Whereas additional semantic content of the sentence or text may contribute to the decision of who is the intended referent of a constituent, for instance, this kind of NP-external relation is never the primary source of information. In most predicative sentences, on the other hand, it is not the features of the

predicative NP and the subject phrase that indicate “identity of reference”, but the sentence predicate usually the verb be”. She suggests testing the potential coreference relation by substituting the predicative verb with some other verb. The substitution should not change the interpretation of the NPs if there is an anaphoric relation involved.

Following the above reasoning, M1 and M2 are not to be annotated as anaphoric; M3 and M4 are to be annotated as anaphoric.

The **Set/Subset relation** is an anaphoric relation where the anaphor refers to a subset of a set of entities or is a superset of a previously mentioned linguistic expression in the discourse. For example, the relation between M1 and M2 is set/subset.

> (M1 Three boys) came in the room. (M2 Two) left.
> (M1 The tumors) have not changed. (M2 Two) are stage 3.

In Set/Subset relations, the interpretation of the constituents in the latter markable depends on previous mentions. Set/subset relations are not to be confused with hyponymic or hyperonymic relations, which do not require the previous or latter markable to interpret.

(added 2/16/2009) In general, annotate Set/Subset and Part/Whole relations only when the coreference is in no doubt. For example, “muscularis propria” could be “part/whole” with any of the named intestinal sites. However, we do not annotate that relation as the coreference is not really clear.

Part/Whole relation is a relation where one discourse referent is a part of another discourse referent.

Example:
Her (M1 arm) was scarred but her (M2 hand) was not.

This relation is different from the Set/Subset relation in that: entities that “Part” is referring to is not of the same type as those that “Whole” is referring to. However, in a Set/Subset relation, both “Set” and “Subset” refer to entities of the same type. For instance, in the sentence “The engine of the car was broken”, “The engine” and “the car” form a Part/Whole relation since they are not of the same type, i.e., an engine is not a type of car. In another sentence “We have replaced two of our six cars”, “two” and “our six cars” form a Set/Subset relation as both are of the same type: car, i.e. the two replaced cars are a kind of car. Note, however, that this simple “is a type/kind of” test is not *sufficient* to determine the relation. For example, “There were already 30 students in the room when another three came in.” Apparently, “30 students” and “another three” are not in a Set/Subset relation even though the three students are a kind of student, because the “another three” students are not in the set of the “30 students.”

The **Other** relation category is a catch-all category for relations different than Identity, Part/Whole and Set/subset such as contrastive “one”.

Example:
(M1 A big room) and (M2 a small one).

Annotation of repeated paragraphs
Definition of repeated paragraphs: The information in the two paragraphs is repeated almost verbatim.

If there are repeating paragraphs, annotate the pairs within each paragraph following the guidelines for coreferring markables. In addition, annotate the markables as pairs across the paragraphs that indicate identity.

Example:
#8 Health-maintenance
Colorectal cancer screening in 1994. Repeat in 1998. (M1 Immunizations) received locally. (M2 Flu vaccine) in October. (M3 Pneumonia) 5-6 years ago. Patient will check with his local physician

[start section id="20106"]

Dry skin. Pruritus on extremities and back. PND with occasional cough. (M4 Immunizations) received locally with (M5 pneumovax) approximately 5-6 years ago, had (M6 a flu vaccine) in October. Disrupted snoring. Questionable sleep apnea

Pairs:
M1 and M2 (set/subset);
M1 and M3 (set/subset);
M4 and M5 (set/subset);
M4 and M6 (set/subset);
M1 and M4 (identity);
M2 and M6 (identity);
M3 and M5 (identity);

Chains:
Flu chain (M1 and M2; M1 and M4; M2 and M6; M4 and M6)
Pneumonia chain (M1 and M3, M1 and M4, M3 and M5; M4 and M4)
Immunization chain: (M1 and M4)

7 Chains

A set of markables that are anaphoric constitute a chain. The grounding instance for an anaphoric chain is the first markable.
Example:
I discussed my clinical expression at length with (M1 Mr. Smith) and (M2 his) wife. I have recommended (M3 he) apply DesOwen lotion b.i.d. prn.

Markable expressions in the chain: M1, M2 and M3
Pairs in the chain: (1) M1 && M2, an (2) M1 && M3

7.1 Attributes

7.1.1 Pairs

Include all pairs that are relevant to the specific chain

7.1.2 Chain Relation Type

The CHAIN RELATION TYPE attribute indicates the relation between all markables in the chain. The types for chains are:

- Same as participating pairs
- Mixed which indicates that the participating pairs have different types, e.g. one pair in the chain might have IDENTITY coreference type, while other pair in the chain might have SET/SUBSET anaphoric type.

Example:
The patient did undergo endoscopy by (M1 the Gastroenterology Service) while in the Emergency Department. As per (M2 their) procedure note, (M3 they) were unable to withdraw the food bolus.

Markables: M1, M2, M3
Pairs: M2 and M1 of relation set/subset; M3 and M1 of relation set/subset
Chain relation: same

References:

Bagga, Amit. 1998. Evaluation of coreferences and coreference resolution system. In Proc. of the First Language Resource and Evaluation Conference. May 1998.

Bodenrieder, Olivier and McCray, Alexa. 2003. Exploring semantic groups through visual approaches. Journal of Biomedical Informatics 36 (2203) 414-432.

Borthen, Kaja. 2004. Predicative NPs and the annotation of reference chains. COLING-2004.

Borthen, Kaja. Last accessed June 20, 2008. Annotation scheme for BREDT. http://bredt.uib.no/publikationer/bredt_annotation.pdf

Castano J., Zhang J. and Pustejovsky J. 2002. Anaphora resolution in biomedical literature. International Symposium on Reference Resolution, 2002.

MUC-7. 1997. MUC-7 Coreference task definition. http://www-nlpir.nist.gov/related_projects/muc/proceedings/co_task.html

Olsson, Fredrik. 2004. A survey of machine learning for reference resolution in textual discourse. SICS Technical Report T2004:02. Swedish Institute of Computer Science ISSN 1100-3154.

Yang, Xiaofeng; Su, Jian; Lang, Jun; Tan, Chew Lim; Liu, Ting; Li, Sheng. 2008. An entity mention model for coreference resolution with inductive logic programming. Proc. of ACL-08: HLT, pages 843-851, Columbus, Ohio, USA, June 2008.