UMLS Entity Annotation Guidelines

for use with the MiPACQ project
Dann Albright, Will Styler

January 2012

Contents

1. Introduction

This document serves as a guide to completing UMLS entity annotation for the MiPACQ project. These guidelines may be used for other projects, but it is important to bear in mind that numerous changes were made to the original UMLS schema to make it better suited to the needs of MiPACQ project, and these changes may not be suitable for the needs of other tasks.

1.1 What Is UMLS?

The Unified Medical Language System (UMLS) is an ontology that was originally created as a semantic tool for use in the creation of technological tools for sharing biomedical information between computer systems. It contains two types of vocabularies: entities and relations (we are using only the entities for this project). We have adapted the UMLS as an annotation schema for use with various types of biomedical texts: clinical notes, pathology reports, encyclopedia entries, and others.

1.2 Annotation Representation Conventions

In this document, the following conventions will be used when giving examples of annotations:

- entities will be enclosed in square brackets, e.g. [leukemia];
- attributes will be enclosed in parentheses, and placed *before* the argument that they apply to, e.g. [chest] (negated) location_of [pain];
- sample text will be written in fixed-width font and centered on a new line, e.g.

  ```
  The patient reports difficulty breathing on waking.
  ```

- example annotations will be placed immediately below sample text.

1.3 General Tips

UMLS annotation is not easy. This is especially true because no one else has undertaken a project like this before. Because of this, we don't have much to compare our work to. However, if you're on this project, you've proven that you have at least some medical knowledge, and you've received training that should enable you to determine how to make the best annotations. Remember: go with your instincts. If you're not completely sure what something, but you have a pretty strong feeling, go with it.

If you're struggling with a particular sentence, e-mail the sentence to your supervisor. He or she can forward that sentence to our medical consultant, an experienced biomedical professional, who will be able to provide insight into the terminology or sentence construction that's giving you trouble.

Use outside resources. See section 6 for a list of resources that you may find useful if you're struggling. Medical texts can be tough to work through, but there are a lot of ways to find help.

2. Overview of Annotation Process

2.1 Note, Set, and Corpus Structure

The MiPACQ medical corpus, to enhance the ease of annotation, has been broken into several smaller sub-corpora. Each sub-corpus consists of several sets, each of which contains a number of text files. Each text file represents a single clinical note, pathology note, or encyclopedia entry. Annotation will be completed set-by-set, and sub-corpus-by-sub-corpus.

2.2 Entity Annotation

Entity annotation is completed by the MiPACQ annotators on all notes. Many notes come with some entities pre-annotated by cTAKES, but all annotators should make a pass over the note to ensure that the cTAKES annotations adhere to project guidelines and to annotate any entities that have been missed. In other notes, there will be no pre-annotations, and all entities must be created by MiPACQ annotators.

In general, use full noun phrases for entity annotations.

2.2.1 Double Entity Annotation

It is important to note that at times, more than one entity annotation will be required on a given string. This occurs most often when a relation can be created between two entities, but the combination of these two entities also constitutes an entity that is required for a relation. When in doubt, annotate entities with an excruciating level of detail.

2.3 Relation Annotation

Once entity annotation is complete, annotators will identify and annotate all intra-sentential relations in a note. All relations must have two entities as its arguments, and it can be difficult to predict which relations will be used during annotation, so it will often be necessary to add entity annotations during the relation annotation phase.

2.5 Adjudication

After the completion of steps 2.2–2.5, all sets that have been double-annotated will be submitted to a medical expert for adjudication. The expert will look at discrepancies between the notes and choose one annotation in each case to become the gold standard. After this step, sets are completely finished, and the gold-standard-annotated data will be submitted to the machine learning team.

3. Entity Annotation

Entities are the “things” of medical notes. They are most often nouns or noun phrases, but if there is a long, complicated string, an entity can span an entire sentence. When annotating entities, try to use the most specific annotation that you can. Many times it will be very difficult to determine what a particular word refers to. In this case, you may e-mail your supervisor, who will consult the medical consultant. Other times, it’s best to use a more vague entity, like “Idea_or_concept,” or a higher-level annotation such as “Procedures.”

Please note that we are not using all subcategories in the schema. In many cases, we don’t need the level of detail that the subcategories provide, and in others, we just don’t have the medical knowledge to use them correctly.

3.1 Activities_and_behaviour

An operation or series of operations that an organism or machine carries out or participates in. Any of the psycho-social activities of humans or animals that can be observed directly by others or can be made systematically observable by the use of special strategies.

3.2 Anatomy

A normal or pathological part of the anatomy or structural organization of an organism.

3.3 Chemicals_and_drugs

Compounds or substances of definite molecular composition.

3.4 Concepts_and_ideas

A broad type for grouping abstract entities or concepts.

3.4.1 Idea_or_concept

An abstract concept, such as a social, religious or philosophical concept.

3.4.2 Intellectual_product

A conceptual entity resulting from human endeavor; usually information created by humans for some purpose.

3.4.3 Qualitative_concepts

A concept which is an assessment of some quality, rather than a direct measurement.

3.4.4 Quantitative_concepts

A concept which involves the dimensions, quantity or capacity of something using some unit of measure, or which involves the quantitative comparison of entities.

3.4.5 Spacial_concept

A location, region, or space, generally having definite boundaries.

3.4.6 Temporal_concept

A concept which pertains to time or duration.

3.5 Devices

A manufactured object used primarily in the diagnosis, treatment, or prevention of physiologic or anatomic disorders.

3.6 Disorders

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.

3.7 Gene_and_molecular_sequences

A specific sequence, or in the case of the genome the complete sequence, of nucleotides along a molecule of DNA or RNA (in the case of some viruses) which represent the functional units of heredity.

3.8 Geographic_areas

A geographic location, generally having definite boundaries.

3.9 Living_beings

A broad type for grouping organisms.

3.9.1 Age_group

An individual or individuals classified according to their age.

3.9.4 Bacterium

A small, typically one-celled, prokaryotic micro-organism.

3.9.5 Family_group

An individual or individuals classified according to their family relationships or relative position in the family unit.

3.9.6 Fungus

A eukaryotic organism characterized by the absence of chlorophyll and the presence of a rigid cell wall. Included here are both slime molds and true fungi such as yeasts, molds, mildews, and mushrooms.

3.9.7 Human

Modern man, the only remaining species of the Homo genus.

3.9.8 Invertebrate

An animal with no vertebral column.

3.9.9 Organism

Generally, a living individual, including all plants and animals.

3.9.10 Patient_or_disabled_group

An individual or individuals classified according to a disability, disease, condition or treatment.

3.9.12 Population_group

An indivdual or individuals classified according to their sex, racial origin, religion, common place of living, financial or social status, or some other cultural or behavioral attribute.

3.9.13 Professional_or_occupational_group

An individual or individuals classified according to their vocation.

3.9.14 Rickettsia_or_Chlamydia

Any organism belonging to the species Rickettsia or Chlamydia.

3.9.15 Virus

An organism consisting of a core of a single nucleic acid enclosed in a protective coat of protein. A virus may replicate only inside a host living cell. A virus exhibits some but not all of the usual characteristics of living things.

3.10 Objects

A broad type for grouping both natural and man-made artifacts.

3.10.1 Food

Any substance generally containing nutrients, such as carbohydrates, proteins, and fats, that can be ingested by a living organism and metabolized into energy and body tissue. Some foods are naturally occurring, others are either partially or entirely made by humans.

3.10.2 Manufactured_object

A physical object made by human beings.

3.10.3 Physical_object

An object perceptible to the sense of vision or touch.

3.10.4 Substance

A material with definite or fairly definite chemical composition.

3.11 Occupations

A vocation, academic discipline, or field of study, or a subpart of an occupation or discipline.

3.12 Organizations

A broad type for grouping groups of people.

3.12.1 Health_care_related_organization

An established organization which carries out specific functions related to health care delivery or research in the life sciences.

3.12.2 Organization

The result of uniting for a common purpose or function. The continued existence of an organization is not dependent on any of its members, its location, or particular facility. Components or subparts of organizations are also included here. Although the names of organizations are sometimes used to refer to the buildings in which they reside, they are not inherently physical in nature.

3.13 Person

A specific, nameable person.

3.14 Phenomena

A broad type for grouping entities and processes that occur naturally or because of humans.

3.14.1 Biologic_function

A state, activity or process of the body or one of its systems or parts.

3.14.2 Environmental_effect_of_humans

A change in the natural environment that is a result of the activities of human beings.

3.14.3 Human_caused_phenomenon_or_process

A phenomenon or process that is a result of the activities of human beings.

3.14.4 Laboratory_or_test_result

The outcome of a specific test to measure an attribute or to determine the presence, absence, or degree of a condition.

3.14.5 Natural_phenomenon_or_process

A phenomenon or process that occurs irrespective of the activities of human beings.

3.14.6 Phenomenon_or_process

A process or state which occurs naturally or as a result of an activity.

3.15 Physiology

A broad type for grouping functions and attributes of the body.

3.15.1 Cell_function

A physiologic function inherent to cells or cell components.

3.15.2 Clinical_attribute

An observable or measurable property or state of an organism of clinical interest.

3.15.3 Mental_process

A physiologic function involving the mind or cognitive processing.

3.15.4 Molecular_function

A physiologic function occurring at the molecular level.

3.15.5 Organ_or_tissue_function

A physiologic function of a particular organ, organ system, or tissue.

3.15.6 Organism_attribute

A property of the organism or its major parts.

3.15.7 Organism_function

A physiologic function of the organism as a whole, of multiple organ systems, or of multiple organs or tissues.

3.15.8 Physiological_function

A normal process, activity, or state of the body.

3.16 Procedures

A broad type for grouping clinical activities.

3.16.1 Diagnostic_procedure

A 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.

3.16.2 Educational_activity

An activity related to the organization and provision of education.

3.16.3 Health_care_activity

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

3.16.4 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.

3.16.5 Research_activity

An activity carried out as part of research or experimentation.

3.16.6 Therapeutic_or_preventive_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.

3.17 Sign_Symptom

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.

4. Negation and Attributes

All entities can be negated, and have several attributes that may be associated with them. Negation and attributes are covered in their own sections below.

5.1 Negation

Negation is indicated by changing the “Negation” attribute to “true.” It’s important to note that “true” does not indicate that the annotation is true—it indicates that the *negation* is true. So if you have annotated [chest pain] as a Sign_Symptom and marked the Negation attribute as true, this will indicate the lack of chest pain.

5.2 Entity Attributes

5.2.1 Status

The Status field has three possible values. “Possible” indicates that there is doubt as to whether the entity exists. “HistoryOf” indicates that a given patient has a history of the annotated entity; this has no bearing on whether the entity is present at the moment. It simply states that it has been present in the past. “FamilyHistoryOf” is similar, though it indicates that the patient has a family history of a particular entity. Again, this does not imply that it is present at the time of the note. The Status field defaults to “none.”