Filtering the UMLS® Metathesaurus® for MetaMap

2010 Edition

Francois-Michel Lang and Alan R. Aronson

July 20, 2010

1. Overview

The MetaMap program's purpose is to discover the Metathesaurus concepts referred to in arbitrary text. A given Metathesaurus concept can have many alternative names (Metathesaurus strings) which originate in the many source vocabularies included in the Metathesaurus. As the number of strings has grown over the years, MetaMap's performance has suffered. In the 2010AA version of the Metathesaurus, for example, the Metathesaurus includes 5,394,495 English strings, 5,338,590 (98.96%) of them distinct, comprising 2,194,659 concepts. There are 2.20% more English strings and 3.51% more concepts than in the 2009AA edition. Many of the strings in the Metathesaurus are of little value to MetaMap for one of four reasons:

1. Some strings are virtually indistinguishable from each other; for efficiency, only one representative of a set of indistinguishable strings is needed.
2. Some strings either represent general, nonmedical concepts, are unnecessarily ambiguous, or have been found to be problematic for some other reason.
3. Some strings have an assigned type in their vocabulary because they have a form (e.g., an idiosyncratic abbreviation) that is highly unlikely to appear in regular text.
4. Some strings, including lengthy descriptions of things such as procedures, health activities or medical devices, are so complicated that it is again unlikely to find them in normal text.

Corresponding to the four classes of strings are four filtering methods for discovering and removing them:

1. lexical filtering,
2. manual filtering,
3. filtering by type, and

4. syntactic filtering.

These methods are discussed in sections 2-5. Then section 6 describes ways to selectively combine the filtering methods to produce a range of alternative views of the Metathesaurus appropriate for various purposes.

2. Lexical Filtering

Lexical filtering is the most benign type of filtering and consists of removing strings for a concept which are effectively the same as another string for the concept. Properties which can make strings effectively the same are:

- case variation;
- hyphen variation; and
- possessives.
- syntactic uninversion;
- *NOS* variation;
- non-essential parentheticals;

Lexical filtering is accomplished by normalizing all strings for a given concept and removing all but one string for each set of strings that normalize to the same thing.

2.1 Case variation

Two strings which differ from each only because of case variation normally refer to the same thing. For example, the concept ‘Abdomen’ has strings “abdomen” and “ABDOMEN*”* in addition to “Abdomen” which differ from each other only by case. Similarly, the concept ‘beta-Alanine’ has strings, “beta Alanine”, “beta alanine” and “BETA ALANINE”, which differ from each other only by case. Note, however, that case *does* matter for some aspects of text processing. Text containing the pronoun *us* is not referring to the acronym *US* for the *United States*; and the verb *aids* does not refer to the disease *AIDS*. Similarly, the case variation in the first three letters of the concepts ‘CDE genotype’, ‘CDe genotype’, ..., and ‘cde genotype’ is significant. Despite these observations, case almost never matters within the limited context of the set of all strings for a given concept.

2.2 Hyphen variation

As with case variation, the presence of a hyphen rather than a space normally means little especially in the context of all strings for a given concept. For example, the concept “1,4-alpha-Glucan Branching Enzyme” used in the last section has a variant “1,4 alpha Glucan Branching Enzyme” in which both hyphens have been replaced by spaces.

2.3 Possessives

Alternatives such as “Down’s Syndrome” and “Down Syndrome” or “American Nurses’ Association” and “American Nurses Association” differ only by a possessive.

2.4 Syntactic uninversion

Inversion refers to the practice of inverting words of a term and inserting a comma to signal the inversion. It is normally done to index the original term under each of its important words and thereby make it more accessible. Inverted forms of a term, however, are not useful for processing text since inverted forms rarely appear in text. The concept "1,4-alpha-Glucan Branching Enzyme" has some interesting inversions. It has a synonym "Branching Enzyme" with inversion "Enzyme, Branching", and it also has a synonym "Starch Branching Enzyme" with two inversions, "Branching Enzyme, Starch" and "Enzyme, Starch Branching". The process of uninversion simply undoes inversion, i.e., it searches for a comma followed by a space, inverts the term at that point and removes the comma and space. Syntactic uninversion is just uninversion which is inhibited if the term contains a preposition or conjunction. This prevents terms such as "Biological Phenomena, Cell Phenomena, and Immunity" or "Legal blindness, as defined in U.S.A." from being incorrectly uninverted. Note that the concept "1,4-alpha-Glucan Branching Enzyme" mentioned earlier is also not uninverted because the comma within it is not followed by a space; such embedded commas do not call for uninversion.

2.5 *NOS* variation

Many of the Metathesaurus vocabularies incorporate the acronym *NOS* (*Not Otherwise Specified*) into their terms. Examples include "Abdomen, NOS" and "X-RAY NEC AND NOS". As with case variation, the presence of *NOS* (except when also accompanied by *NEC*) does not generally have a significant effect on the meaning of the term. The argument for ignoring *NOS* variation is not as strong as that for case variation, but it still seems reasonable for most text processing.

2.6 Non-essential parentheticals

Non-essential parentheticals are parenthetical expressions within a string which provide meta information about the string. As such they are not useful for text processing. Non-essential parentheticals can occur at the left or right end of a string and can be delimited by either parentheses or brackets. For example the concept "Anemia, Hemolytic" has synonyms "[X]Haemolytic anaemias" and "[X]Hemolytic anemias" both of which contain the left parenthetical *[X]*. Previous editions of the Metathesaurus only contained right parentheticals which seemed to be relatively well-behaved in the sense that a string without the parenthetical was almost always present in the set of strings for a given concept. Thus, "Drug Toxicity (Non MeSH)" had a string "Drug Toxicity". Now right parentheticals are much less well-behaved and only a few left parentheticals can be reliably removed without altering the string's meaning. These left parentheticals come from the Read Codes (and also SNOMEDCT): *[X]*, *[V]*, *[D]*, *[M]*, *[EDTA]*, *[SO]* and *[Q]*. These are the only parentheticals declared to be non-essential and removed from strings. The problem of detecting non-essential parentheticals has changed as the Metathesaurus has matured. The current practice of removing the few left parentheticals listed above is by no means adequate. The problem requires further analysis.

3. Manual Filtering

A number of Metathesaurus strings are problematic for various reasons. We have (somewhat arbitrarily) decided that we do not want to map to them. There are 16,525 such strings, 16.91% ***fewer*** than in 2009:

- **Unnecessarily ambiguous terms** [10,571 occurrences]
  'Other' for 'Other location of complaint'
  'Protocols' for 'Protocols: Urinary Elimination'
  …
- **Contextual terms**, i.e., terms whose meaning can only be understood within the context of their vocabulary [5,003 occurrences]
  All terms containing "*NEC"* or an expanded form
- **Brand names**, i.e., short forms of terms containing "*brand"* [260 occurrences]
- **Enzyme Commission (EC) numbers** beginning "*EC <integer>*." [208 occurrences]
- **Numbers** (e.g., '2', '+1', '-4', '98.734', '50000') [232 occurrences];
- **Single alphabetic strings** (e.g., 'a', 'A', 'b', 'B') [211 occurrences];
- **Special cases** [40 occurrences]
  'Periods' and 'Period' for 'Menstruation' (C0025344)
  'Clap' and 'CLAP' for 'Gonorrhea' (C0018081)
  'BRA' for 'Brain' (C0006104)
- …

Before describing the types of manual filtering listed above, we note that the Metathesaurus staff marks some terms as suppressible synonyms because they are thought to be inappropriate for any use. These are terms that, for one reason or another, do not adequately describe the concept that contains them. They are given a Term Status (TS) of lowercase s (or p) and are discussed in the annual editions of *Ambiguity in the UMLS Metathesaurus*.

The first type of manual filtering consists of **unnecessarily ambiguous terms** which are determined annually by a manual review process. **Contextual terms** are actually a specific kind of ambiguous term. They are identifiable by the presence of *NEC* or one of its expansions (e.g., not elsewhere classified). **Brand names** are problematic because they often consist of a common word (e.g., *Cold*) which almost never has the brand name meaning in biomedical text. **Enzyme Commision (EC) numbers** are highly ambiguous and are only sporadically represented in the Metathesaurus. Although a few **numbers** correspond to biomedical entities ("98.734" has semantic types '*Steroid*' and '*Pharmacologic Substance*'), they generally have semantic types '*Quantitative Concept*' or '*Intellectual Product*'. Similarly, the **single alphabetics** often mean the letter itself (the concept for "a" is "Lower case ay") and have semantic type '*Intellectual Product*' (several single alphabetics, however are biomedical: "B" has concept "Boron" with semantic type '*Element, Ion, or Isotope*'). A final class of **special cases** includes the string "Periods" for "Menstruation" and "BRA" for "Brain". Both of these are problematic because they are ambiguous with other concepts which occur far more frequently in biomedical text.

4. Filtering by Type

For 2009AA, a study was done to evaluate the efficacy of filtering by Term Type and the decision was to not do this for 2009AA. The original study was designed around the idea of trying to automate the determination of "Good", "Bad", and "Ugly" Term Types and it became clear that in the wholesale removal of the Term Types, we were inadvertently removing some valid concepts. Based on this observation, we decided to forego the Term Type filtering for 2009AA and to track the data over the course of the year to see how MetaMap behaves. Filtering by Type was the major differentiation between the Moderate and Relaxed Models, so the fact that we are no longer doing this filtering along with the fact that we saw very little use of the Moderate Model, we have decided to forego creating the Moderate Model.

5. Syntactic Filtering

The final kind of filtering considered here is based on a high-level syntactic parse of the Metathesaurus strings. Since normal MetaMap processing involves mapping the simple noun phrases found in text, it is highly unlikely that a complex Metathesaurus string will be part of a good mapping. For example, the concept "Accident caused by caustic and corrosive substances" has high-level syntactic analysis [[head],[verb],[prep,head],[conj],[mod,head]] which contains seven syntactic units (head, verb, etc.) broken into five simple phrases ([head], [verb], etc.) Any text which resembles the concept will be broken up into several phrases each of which is processed separately. Thus, the text might map to constituent concepts (such as "Accident"); but the entire text will not map to the full concept. The strictest form of syntactic filtering, then, would be to filter out any string consisting of more than one simple phrase. However some tractable strings with more than one simple phrase are not filtered out. As of 1999, for example, strings containing *of* such as "Acute necrosis of liver" and "Radical resection of tumor of soft tissue of leg area", which consist of a simple phrase followed by one or more *of* prepositional phrases, have not been excluded in syntactic filtering because of their tractability. In 2001 this condition was relaxed further to include phrases consisting of a simple phrase followed by any prepositional phrase followed by zero or more *of* prepositional phrases. An example of such a phrase is "Other operations on vessels of heart".

6. Filtered Metathesaurus Models

The filtering described in the previous sections can be selectively applied to provide different views of the Metathesaurus. Three such models are

- Strict Model: All forms of filtering, lexical, manual, and syntactic, are applied. This view is most appropriate for semantic processing where the highest level of accuracy is needed. The Strict Model consists of 2,427,017 (44.99%) of the 5,394,495 English Metathesaurus strings;
- Moderate Model: Lexical, manual, and type-based filtering, but not syntactic filtering, are used. This view is appropriate for term processing where input text should not be divided into simple phrases but considered as a whole. ***The Moderate Model was disconstinued beginning with the 2009AA release because we believe it was very rarely used, and we no longer performed***

Term-Type Filtering, which was the major difference between the Moderate and Relaxed Models; and

- Relaxed Model: Only lexical and manual filtering are performed. This provides access to virtually all Metathesaurus strings and is appropriate for browsing. The Relaxed Model consists of 4,998,184 (92.65%) of the 5,394,495 English Metathesaurus strings.

Note that before 2009, in order to have all Metathesaurus concepts represented in each model, the preferred name of a concept was retained when all of its strings would otherwise be filtered out. We no longer do this; so if all of a concept's strings are filtered out, the concept disappears.