MetaMap 2013 Release Notes

August 26, 2013

MetaMap 2013 includes several interesting user-driven enhancements described below.

1 New Berkeley-DB Table

MetaMap 2013 relies on a new Berkeley-DB table that does not exist in previously-downloaded datasets (2012AA, 2012AB, etc.), so we strongly encourage using the 2013AA data with MetaMap 2013. In order to ensure backward compatibility with previous datasets, however, we have added a version of that table to all pre-2013 datasets, which are available at

`http://metamap.nlm.nih.gov/#OptionalDatasets`

Users needing to run MetaMap2013 with previous datasets should download and re-install them.

2 NegEx Enhancements

Sina Madani of the MD Anderson Cancer Center requested some significant extensions to MetaMap's NegEx processing, which are described next.

These NegEx enhancements required changes to MetaMap's output formats, as described below; consequently user-developed programs that postprocess MetaMap output will probably require modifications.

2.1 Background

MetaMap has for several years included an implementation of Wendy Chapman's NegEx negation-detection algorithm, which is documented at `http://code.google.com/p/negex`. NegEx is invoked from command-line MetaMap simply by using the `--negex` option. Changes to the various forms of MetaMap's NegEx output are highlighted in Section 2.3 below.

MetaMap generates human-readable NegEx output only if NegEx is explicitly invoked; however, NegEx information is automatically included in MetaMap's Prolog-based Machine Output (invoked with `-q`), and XML output (invoked with one of `--XMLf`, `--XMLn`, `--XMLf1`, and `--XMLn1`). Indeed specifying `--negex` when requesting MMO or XML results in a warning message, because the `--negex` option is redundant if MMO or XML output is generated.

2.2 Extensions to NegEx: User-Specified Semantic Types

These extensions are experimental. We have not tested NegEx's behavior with Semantic Types beyond those specified in original NegEx implementation; consequently, some NegEx triggers may work well with other Semantic Types, but others perhaps not. A list of all Semantic Types showing the short name, unique identifier, and long name is available at
`http://metamap.nlm.nih.gov/SemanticTypes_2013AA.txt`.

MetaMap's NegEx by default identifies concepts of only the following Semantic Types, which were specified by Wendy Chapman in the original NegEx implementation:

acab	Acquired Abnormality
anab	Anatomical Abnormality
biof	Biologic Function
cgab	Congenital Abnormality
comd	Cell or Molecular Dysfunction
dsyn	Disease or Syndrome
emod	Experimental Model of Disease
fndg	Finding
inpo	Injury or Poisoning
lbtr	Laboratory or Test Result
menp	Mental Process
mobd	Mental or Behavioral Dysfunction
neop	Neoplastic Process
patf	Pathologic Function
phsf	Physiologic Function
sosy	Sign or Symptom

We have added the following MetaMap command-line options to enable users to specify the Semantic Types used by MetaMap's NegEx implementation:

1. `negex_st_add` takes as a mandatory argument a comma-separated list of SemTypes (as do the existing command-line options `exclude_sts` and `restrict_to_sts`) to be *added* to the default NegEx SemTypes. E.g., `--negex_st_add amph,anim,bird,fish,mamm,rept`.
2. `negex_st_del` takes a list of SemTypes to be *deleted* from the default NegEx SemTypes.
3. `negex_st_set` takes a list of SemTypes to *replace* the default NegEx SemTypes.

Some other features of this new functionality:

- The `--negex` option can still be used without any of the three new options described above, in which case MetaMap will simply use the default SemTypes shown in the table above.
- If any of the three new options is specified, it is not necessary to specify `--negex`, although doing so is permissible and will generate no warning.

- If either `all` or `ALL` is specified as a meta-SemType with `negex_st_add` or `negex_st_set`, *all* SemTypes will be used instead of the default NegEx SemTypes. A fatal error is generated if either `all` or `ALL` is used with `negex_st_del`.
- `negex_st_add` and `negex_st_del` can be used together in order to both add and delete SemTypes. If both are used, the specified SemType additions are done first, followed by the specified SemType deletions, regardless of the order in which the two options appear on the command line.
- A fatal error is generated if `negex_st_set` is used together with either `negex_st_add` or `negex_st_del`.
- A fatal error is generated if any one of the three new options is specified with an invalid SemType (just as with `exclude_sts` and `restrict_to_sts`), or with no SemType argument at all.
- A warning is generated if `negex_st_add` is called with a valid SemType that is already one of the default NegEx SemTypes, or `negex_st_del` is called with a valid SemType that is not one of the default NegEx SemTypes, because neither of these actions has any effect.

2.3 NegEx Output Changes

We have also modified MetaMap's output formats (human-readable, Prolog Machine Output, and XML) to reflect more detailed negation information.

We now describe the changes to MetaMap's NegEx output, using as input text

`No pneumonia.  Checked for infiltrates.`

Changes in each output format are highlighted in **red**.

2.3.1 Human-Readable Candidates and Mappings Output

Human-readable candidates and mappings output now includes "N" between the candidate score and the Metathesaurus string for negated candidates (highlighted in **red** below).

```
Phrase: "no pneumonia."
Meta Candidates (Total=5; Excluded=4; Pruned=0; Remaining=1)
  1000 N Pneumonia (Pneumonia) [Disease or Syndrome]
   907 E LUNGS (Lung) [Body Part, Organ, or Organ Component]
   907 E Lung (Entire lung) [Body Part, Organ, or Organ Component]
   893 E Pulmonary (Pulmonary:-:Point in time:^Patient:-) [Clinical Attribute]
   893 E Pulmonary (Pulmonary (qualifier value)) [Qualitative Concept]
Meta Mapping (1000):
  1000 N Pneumonia (Pneumonia) [Disease or Syndrome]
Processing 00000000.tx.2: Checked for infiltrates
```

```
Phrase: "Checked"
Meta Candidates (Total=1; Excluded=0; Pruned=0; Remaining=1)
  1000   Checked (Checking (action)) [Activity]
Meta Mapping (1000):
  1000   Checked (Checking (action)) [Activity]

Phrase: "for infiltrates"
Meta Candidates (Total=3; Excluded=2; Pruned=0; Remaining=1)
  1000 N INFILTRATES (Infiltration) [Pathologic Function]
   966 E Infiltrate (Specimen Source Codes - Infiltrate) [Intellectual Product]
   966 E Infiltrate (Administration Method - Infiltrate) [Functional Concept]
Meta Mapping (1000):
  1000 N INFILTRATES (Infiltration) [Pathologic Function]
```

2.3.2 Candidates and Mappings Prolog Machine Output (MMO)

MMO for candidates (including candidates appearing in the mappings) now includes an additional final argument which is set to 1 if the concept is negated and 0 if not. The (pretty-printed and abbreviated) candidates Machine Output for for the above text is below, again with the new 1/0 field highlighted in **red**.

```
candidates(5,4,0,1,
   [ev(-1000,'C0032285','Pneumonia','Pneumonia',[pneumonia],[dsyn],
       [[[1,1],[1,1],0]],yes,no,['AOD','CHV','COSTAR'],[3/9],0,1),
    ev(-907,'C0024109','LUNGS','Lung',[lungs],[bpoc],
       [[[1,1],[1,1],5]],yes,no,['AOD','CHV','CSP','FMA'],[3/9],1,0),
    ev(-907,'C1278908','Lung','Entire lung',[lung],[bpoc],
       [[[1,1],[1,1],5]],yes,no,['MTH','SNOMEDCT'],[3/9],1,0),
    ev(-893,'C2707265','Pulmonary','Pulmonary',[pulmonary],[clna],
       [[[1,1],[1,1],7]],yes,no,['LNC_MDS20','MTH'],[3/9],1,0),
    ev(-893,'C2709248','Pulmonary','Pulmonary',[pulmonary],[qlco],
       [[[1,1],[1,1],7]],yes,no,['AIR','LNC','MTH'],[3/9],1,0)
   ]).
```

2.3.3 XML NegEx Output

XML output always includes NegEx output. The NegEx component XML output has not changed; however, XML output for candidates (including candidates appearing in the mappings) includes

an additional final tag *Negated* which takes on values 1 (meaning that the concept is negated) or 0 (concept is not negated). The (abbreviated) XML output for the first candidate generated by the above text is below, again with the new tag in **red**.

```
<Candidate>
 <CandidateScore>-1000</CandidateScore>
 <CandidateCUI>C0032285</CandidateCUI>
 <CandidateMatched>Pneumonia</CandidateMatched>
 <CandidatePreferred>Pneumonia</CandidatePreferred>
 <MatchedWords Count="1">
  <MatchedWord>pneumonia</MatchedWord>
 </MatchedWords>
 <SemTypes Count="1">
  <SemType>dsyn</SemType>
 </SemTypes>
 <MatchMaps Count="1">
  <MatchMap>
    . . . MatchMap XML . . .
  </MatchMap>
 </MatchMaps>
 <IsHead>yes</IsHead>
 <IsOverMatch>no</IsOverMatch>
 <Sources Count="18">
    . . . Sources XML . . .
 </Sources>
 <ConceptPIs Count="1">
  <ConceptPI>
   <StartPos>3</StartPos>
   <Length>9</Length>
  </ConceptPI>
 </ConceptPIs>
 <Status>0</Status>
 <Negated>1</Negated>
</Candidate>
```

3 Blanklines Feature

Steven Bedrick of OHSU requested the ability to specify on the command line the number of empty or whitespace-only lines required to end a citation. By default, MetaMap considers an input record or citation to end when it encounters a single empty line or a line containing only whitespace.

This new feature can be invoked as `--blanklines N`, where `N` must be a positive integer, and will result in MetaMap's not ending a citation until `N` blank/whitespace lines have been read. We expect the `--blanklines` option will be particularly useful for analyzing clinical text, which frequently includes multiple blank/whitespace lines in the middle of reports, e.g.,

EXAM

PA and lateral chest radiograph

Three views of the left shoulder
All exams XXXX, XXXX

COMPARISON
None

INDICATION
Syncope, XXXX from ladder

FINDINGS
See impression

IMPRESSION

Chest
Three total images. The heart size is within normal limits.
Mildly tortuous thoracic aorta. No abnormal mediastinal widening
is appreciated.Normal pulmonary vascularity. No pleural effusion
or pneumothorax. There is an S-shaped curvature of the thoracolumbar
spine and a mild kyphosis at the thoraco lumbar junction without
clear XXXX deformity identified.

Left shoulder
There is a mildly comminuted fracture at the junction of the middle
and lateral thirds of the left clavicle, the distal most fragment is
displaced superiorly approximately 25% bone width.
Glenohumeral alignment appears preserved without dislocation and
no additional acute fractures are seen. There is mild superior
subluxation of the humerus on the glenoid which suggests reflect
chronic rotator XXXX pathology; dysmorphic ossification superolateral
to the humeral head XXXX reflecting calcific tendinitis.

Indiana University collection. Case 41.
Xray Chest PA and Lateral
Marc D Kohli, Marc Rosenman
Affiliation: Indiana University School of Medicine, Regenstrief Institute, Inc.

4 Clarification about Term-Processing Option

It was recently brought to our attention that although the `term_processing` option substantially increases recall when used by itself, certain potentially relevant concepts are still missed when MetaMap uses its default Strict Model because those concepts exist in the Relaxed Model only. Several straightforward examples that MetaMap will recognize only if `--term_processing` is used with the Relaxed Model are the following:

`Echocardiography, Doppler, Color` (C0013521)

`Glucose, Blood, Self-Monitoring` (C0005803)

`Abnormalities of size and form of teeth` (C0000770)

`Aseptic necrosis of head and neck of femur` (C0003977)

5 Clarification about Composite-Phrases Option

A composite phrase consists of

- a noun phrase followed by
- any prepositional phrase, optionally followed by
- one or more prepositional phrases introduced by *of*.

Our canonical composite phrase is *pain on the left side of the chest*, which MetaMap by default processes as three small phrases e.g., [*pain*] [*on the left side*] [*of the chest*].

The `--composite_phrases N` option causes MetaMap to construct longer, composite phrases from the smaller phrases produced by the parser; N is the number of prepositional phrases that can be glommed onto the initial noun phrase. MetaMap users may experience increased recall with `--composite phrases` (e.g., `-Q 2`, `-Q 3`, or even `-Q 4`), because it enables the identification of concepts such as `Left sided chest pain` (C0541828) from the text *pain on the left side of the chest*.

Caveat: Using this option may result in increased processing time, and, with extremely complex text, cause occasional out-of-memory errors. We nonetheless encourage users to experiment with this option; indeed in subsequent versions of MetaMap, the `composite_phrases` option will be on by default, although the option will be dynamically disabled for phrases that could experience long runtimes or out-of-memory errors; moreover, the default option can always be overridden by using `-Q 0`.