MetaMap 2011v2 Release Notes

May 29, 2012

MetaMap2011v2 includes some significant enhancements, most notably revisions to the mapping-construction algorithms that further speed the analysis of text that had remained problematic for MetaMap 2011. This release also includes improvements to the format of all MetaMap output (human readable, Prolog Machine Output, and XML), as well as two new input formats. **The changes to MetaMap's output format will almost certainly require modifications to user-developed programs that postprocess MetaMap output.**

1 Further Algorithmic Improvements

MetaMap first implemented algorithmic improvements such as improved handling of duplicate candidates and candidate pruning in the 2011 release; these improvements are described in the MetaMap 2011 Release notes, available at

`http://metamap.nlm.nih.gov/MM_2011_ReleaseNotes.pdf`.

In the initial implementation of candidate pruning, MetaMap by default pruned the candidate set to 35 candidates, but allowed the user to

- disable candidate pruning altogether by using the `--no_prune` command-line option, or
- set a customized pruning threshold by using the `--prune N` command-line option.

In MetaMap2011v2, we have made the pruning algorithm more sensitive to phrase length and the aggregate phrase coverage of the candidate concepts. For example, consider the input phrase

> *with ultra-high spatial resolution black blood inner volume three-dimensional fast spin echo magnetic resonance imaging*

(PMID 18080213), which contains 16 non-stopwords, and from which MetaMap identifies 24 candidates after temporarily removing duplicates. In order to determine if candidate pruning is necessary, MetaMap computes a phrase-coverage grid, shown in Figure 1 below, in which the candidate concepts are ordered by decreasing order of score. The grid identifies which of the 16 phrase positions are covered by each of the 24 candidate concepts: For example, the first candidate, which consists of the words *three*, *dimensional*, and *imaging*, covers phrase positions 9–16; similarly, the tenth candidate (*high*, *resolution*) covers positions 2–4.

MetaMap then calculates the grid's sparseness, i.e., the percentage of the grid that is covered by at least one candidate: In the grid below, 59 of the 384 (=24*16) phrase positions are covered, or 15.36%. Using the sparseness measure, MetaMap's new candidate-pruning algorithm then executes the following logic:

```
   |123456789|123456|
 1 |        ********|  C0887832  835 [three,dimensional,imaging]
 2 |     **********|  C0243032  808 [vascular,magnetic,resonance,imaging]
 3 |     **********|  C0546783  805 [vascular,imaging]
 4 |             ***|  C0024485  802 [magnetic,resonance,imaging]
 5 |             ** |  C0028580  793 [magnetic,resonance]
 6 |               *|  C0011923  785 [imaging]
 7 |             *  |  C0024488  785 [magnetic]
 8 |              * |  C0231881  785 [resonance]
 9 |     ***        |  C0005850  632 [blood,volume]
10 | ***            |  C1719039  632 [high,resolution]
11 |        **      |  C0450363  626 ['3',dimensional]
12 |    *           |  C0005680  618 [black]
13 |     *          |  C0005767  618 [blood]
14 |          *     |  C0015663  618 [fast]
15 |            *   |  C0058928  618 [echo]
16 |      *         |  C0205102  618 [inner]
17 | *              |  C0205250  618 [high]
18 |        *       |  C0205449  618 [three]
19 |       *        |  C0449468  618 [volume]
20 |*               |  C0728475  618 [ultra]
21 |           *    |  C1415347  618 [spin]
22 |  *             |  C1428114  618 [spatial]
23 |   *            |  C1514893  618 [resolution]
24 |         *      |  C0439534  547 [dimension]
   |123456789|123456|
```

Figure 1: Phrase-Coverage Grid

REPEAT

1. If there are ≥ 45 non-duplicate candidates, unconditionally prune the candidate list to 45.
2. Otherwise, if there are ≥ 24 non-duplicate candidates and the grid's sparseness is $\leq 22\%$, prune the candidate list by one, using the logic described in the MetaMap 2001 Release Notes.

UNTIL no pruning is done in the most recent loop iteration

The specific numbers in the logic above (45, 24, and 22) were established empirically through extensive testing on our Linux machines, and may consequently not be appropriate for all users. Pruning can be explicitly controlled by the user as before via the following two command-line flags:

- the `--no_prune` option suppresses pruning altogether, and
- the `--prune N` option enforces pruning to N candidates.

Users are encouraged to experiment with various pruning levels in order to allow MetaMap to generate the best possible results in a reasonable runtime and without exceeding memory limits. Finally, users wishing to view the candidate-pruning grid can call MetaMap with the `--debug grid` command-line option; however, be aware that doing so may send a large amount of text to standard output.

2 New Output Formats

The form of MetaMap's three output formats (human-readable output, Prolog machine output, and XML) has been enhanced by including additional information about the status of the UMLS candidate concepts identified in the text. We begin with some background about how candidate concepts are handled before mapping construction is undertaken.

2.1 Excluded and Pruned Candidates

MetaMap employs two filtering strategies to eliminate candidate concepts from consideration as participants in final mappings: Exclusion and Pruning.

Exclusion: Candidate exclusion has always been an integral component of MetaMap's mapping-construction algorithm, and was originally described in the MetaMap Mapping Algorithm Technical Document located at

http://skr.nlm.nih.gov/papers/references/mm.mapping.pdf.

A candidate concept is excluded from mapping construction if its score is lower than that of another candidate with the same phrase coverage. For example, from the input *heart attack*, MetaMap identifies the following candidates:[1]

```
Meta Candidates (7):
   1.   1000 Heart attack (Myocardial Infarction) [Disease or Syndrome]
   2.    861 Heart [Body Part, Organ, or Organ Component]
   3.    861 Attack, NOS (Onset of illness) [Finding]
   4.    861 attack (Attack behavior) [Social Behavior]
   5.    861 Heart (Entire heart) [Body Part, Organ, or Organ Component]
   6.    861 Attack (Observation of attack) [Finding]
   7.    827 attacked (Assault) [Injury or Poisoning]
```

In the above output, the phrase coverage of candidate #7 is identical to that of candidates #3, #4, and #6; moreover, candidate #7's score of 827 is lower than that of the other three (861). Candidate #7 will therefore be excluded from mapping construction. Candidate exclusion happens automatically and is not subject to user control.

[1] The numbering of the candidates is provided via the `--number_the_candidates` (`-n`) command-line option.

Pruning: As noted above, the MetaMap 2011 Release Notes explained candidate pruning—in particular MetaMap's default pruning behavior, and how pruning can be controlled by the user; an enhanced pruning strategy is described earlier in this document. Note that candidate pruning is invoked only when the number of candidates exceeds a certain threshold or when explicitly requested by the user.

We next present MetaMap's new output formats, and specifically how they include additional information about the number of excluded and pruned candidates. As noted in the introduction of this document, these changes will probably require modifications to user-developed postprocessing programs.

2.2 New Output Format: Human Readable

Previous versions of MetaMap's output provided no information about how many or which candidates had been excluded or pruned. Beginning in MetaMap 2011v2, however, excluded and pruned candidates will be explicitly identified as such; in addition, the candidates header line preceding the list of candidates will specify how many total candidates were identified, how many were excluded and pruned, and how many candidates remain available for participation in mapping construction. For example, the output for *heart attack* will appear as shown below, showing that candidate #7 was excluded ("E"):

```
Meta Candidates (Total=7; Excluded=1; Pruned=0; Remaining=6)
   1. 1000   Heart attack (Myocardial Infarction) [Disease or Syndrome]
   2.  861   Heart [Body Part, Organ, or Organ Component]
   3.  861   Attack, NOS (Onset of illness) [Finding]
   4.  861   attack (Attack behavior) [Social Behavior]
   5.  861   Heart (Entire heart) [Body Part, Organ, or Organ Component]
   6.  861   Attack (Observation of attack) [Finding]
   7.  827 E [X]Attacked (Assault) [Injury or Poisoning]
```

Pruned candidates will be similarly noted with "P". For example, if a MetaMap user requests pruning to only a single candidate by specifying `--prune 1` on the command line, the candidates portion of human-readable output will be the following:

```
Meta Candidates (Total=7; Excluded=1; Pruned=5; Remaining=1)
   1. 1000   Heart attack (Myocardial Infarction) [Disease or Syndrome]
   2.  861 P Heart [Body Part, Organ, or Organ Component]
   3.  861 P Attack, NOS (Onset of illness) [Finding]
   4.  861 P attack (Attack behavior) [Social Behavior]
   5.  861 P Heart (Entire heart) [Body Part, Organ, or Organ Component]
   6.  861 P Attack (Observation of attack) [Finding]
   7.  827 E [X]Attacked (Assault) [Injury or Poisoning]
```

Users preferring the original output format without the additional exclusion/pruning information and the more informative candidates header line should use the `--silent` command-line option.

2.3 New Output Format: Prolog Machine Output

This section is relevant only for users who postprocess MetaMap's Prolog machine output, which is described at

http://metamap.nlm.nih.gov/MMO_08_Info.html.

In order to maintain consistency with MetaMap's human-readable output, the

`candidates(<ListOfCandidates>)`

term now includes the counts of total, excluded, pruned, and remaining candidates: i.e.,

```
candidates(TotalCandidatesCount, ExcludedCandidateCount,
           PrunedCandidateCount, RemainingCandidates,
           <ListOfCandidates>)
```

In addition, each candidate structure now includes an additional final argument, `Status`:

OLD form of candidates term:

```
ev(Score, CUI, String, PreferredName, Words,
   SemTypes, MatchMap, InvolvesHead, IsOverMatch,
   Sources, PositionalInfo)
```

e.g.,

```
ev(-1000, 'C0027051', 'Heart attack', 'Myocardial Infarction', [heart,attack],
   [dsyn], [[[1,2],[1,2],0]], yes, no,
   ['MEDLINEPLUS', ..., 'MSH'], [0/12])
```

NEW form of candidates term:

```
ev(Score, CUI, String, PreferredName, Words,
   SemTypes, MatchMap, InvolvesHead, IsOverMatch,
   Sources, PositionalInfo, Status)
```

e.g.,

```
ev(-1000, 'C0027051', 'Heart attack', 'Myocardial Infarction', [heart,attack],
   [dsyn], [[[1,2],[1,2],0]], yes, no,
   ['MEDLINEPLUS', ..., 'MSH'], [0/12], 0)
```

The possible values of Status, the 12th and last argument of the candidate structure, are:

- 1 if the candidate was excluded,
- 2 if the candidate was pruned, and
- 0 otherwise, i.e., if the candidate was retained.

The new form of Machine Output is invariable and not subject to user control.

2.4 New Output Format: XML

MetaMap's XML output was originally described in the MetaMap 2008 Release Notes, at

http://metamap.nlm.nih.gov/MM08_Release_Notes.shtml#XML_Output;

an update was described in the MetaMap 2009 Release Notes,

http://metamap.nlm.nih.gov/MM09_Release_Notes.shtml#XML_Generation,

and finally the XML tags were made more mnemonic, as described at

http://metamap.nlm.nih.gov/MM09_v2_XML_Info.shtml.

In MetaMap2011v2, XML output has been modified in order to maintain consistency with human-readable and Prolog Machine Output as described above:

1. The `Candidates` tag now includes attributes specifying the number of total, excluded, pruned, and remaining candidates:
   OLD: `<Candidates Count="7">`
   NEW: `<Candidates Total="7" Excluded="1" Pruned="0" Remaining="6">`.
2. The tag identifying the number of candidates in each mapping has been changed from `Candidates` to `MappingCandidates`; in addition, for consistency with the Candidates tag described immediately above, the `Count` attribute has been changed to `Total`:
   OLD: `<Candidates Count="1">`
   NEW: `<MappingCandidates Total="1">`.
3. Each candidate now contains a `Status` tag, e.g., `<Status>N</Status>`, in which `N` is
   - 1 if the candidate was excluded,
   - 2 if the candidate was pruned, and
   - 0 otherwise, i.e., if the candidate was retained.

The new XML representation of the candidates section of XML output is shown below; note in particular the appearance of `<Status>0</Status>` in the next-to-last line of the XML, immediately above `</Candidate>`.

```
<Candidates Total="7" Excluded="1" Pruned="0" Remaining="6">
 <Candidate>
  <CandidateScore>-1000</CandidateScore>
  <CandidateCUI>C0027051</CandidateCUI>
  <CandidateMatched>Heart attack</CandidateMatched>
  <CandidatePreferred>Myocardial Infarction</CandidatePreferred>
  <MatchedWords Count="2">
   <MatchedWord>heart</MatchedWord>
   <MatchedWord>attack</MatchedWord>
  </MatchedWords>
  <SemTypes Count="1">
   <SemType>dsyn</SemType>
  </SemTypes>
  <MatchMaps Count="1">
   <MatchMap>
    <TextMatchStart>1</TextMatchStart>
    <TextMatchEnd>2</TextMatchEnd>
    <ConcMatchStart>1</ConcMatchStart>
    <ConcMatchEnd>2</ConcMatchEnd>
    <LexVariation>0</LexVariation>
   </MatchMap>
  </MatchMaps>
  <IsHead>yes</IsHead>
  <IsOverMatch>no</IsOverMatch>
  <Sources Count="2">
   <Source>MEDLINEPLUS</Source>
   <Source>MSH</Source>
  </Sources>
  <ConceptPIs Count="1">
   <ConceptPI>
    <StartPos>0</StartPos>
    <Length>12</Length>
   </ConceptPI>
  </ConceptPIs>
  <Status>0</Status>
 </Candidate>
  . . . XML for other candidates omitted . . .
</Candidates>
```

The new form of XML Output is invariable and not subject to user control.

The old DTD for MetaMap2011's XML output is available at

http://metamap.nlm.nih.gov/DTD/MMOtoXML_v4.dtd;

the new DTD for MetaMap2011v2's XML output is available at

http://metamap.nlm.nih.gov/DTD/MMOtoXML_v5.dtd.

3 Single-Line Delimited Output

Our Batch Scheduler Facility, available at

http://skr.nlm.nih.gov/batch-mode/index.shtml,

has long handled two special input formats, namely Single-Line-Delimited Input and Single-Line-Delimited Input with ID. MetaMap 2011v2 now includes the ability to process these two formats:

1. **Single-Line-Delimited Input** (SLDI): Normally, MetaMap assumes that distinct input records, whether individual terms or entire MEDLINE citations, are separated by blank lines. In SLDI mode, each line of the input file is treated as its own input record, and blank lines in the file are ignored. For example, the input text

   ```
   heart attack
   lung cancer
   ```

   is by default treated as a single input record. In SLDI mode, however, `heart attack` and `lung cancer` will be treated as separate input records. To call MetaMap on files formatted as SLDI, simply use the `--sldi` command-line option in addition to any other desired options.

2. **Single-Line-Delimited Input With ID** (SLDI-ID): This input format is an extension of SLDI mode that allows the user to specify input record identifiers analogous to PubMed PMIDs. For example, if SLDI-ID mode is specified, in the input text

   ```
   11234793|heart attack
   Cit-2|lung cancer
   ```

   `11234793` and `Cit-2` will be treated as record identifiers. The record identifier and the actual text should be separated by a pipe symbol ("`|`"); all whitespace immediately before and after the "`|`" is ignored. To call MetaMap on files formatted as SLDI with ID, simply use the `--sldiID` command-line option in addition to any other desired options.

The intent of both these input formats is obviously to facilitate the analysis short chunks of text, and not entire paragraphs. When using these formats, it might be advisable to run MetaMap with the `term_processing` (`-z`) option, which is described at

http://metamap.nlm.nih.gov/MM11_Usage.shtml#term_processing.

4 New Phrase-Breaking Characters

Previous versions of MetaMap considered the following characters

```
:  (  )  ;  <  >  =  *
```

as phrase-breaking characters. We have now added "[" and "]" to the list of phrase-breaking characters. The motivation for this change is text such as

```
Education needed:
   [x]Home care after Gyn surgery
   [ ]Tube feeding
   [ ]Drain care type
   [ ]Central Venous Catheter care
   [ ]Wound
   [ ]Other

Multidisciplinary consults completed:
   [x]Case management
   [ ]Nutrition
   [x]Occupational Therapy
   [x]Physical Therapy
   [ ]Social worker
   [ ]TPN team
   [ ]Wound ostomy
   [ ]Other
```

which would otherwise not be broken up into appropriate small phrases.

5 Retired Command-Line options

The following MetaMap command-line options, which were marked as **DEPRECATED** in the MetaMap 2011 Usage document

http://metamap.nlm.nih.gov/MM11_Usage.shtml

have been retired and are no longer available:

- `--allow_duplicate_concept_names` (short form `-U`),
- `--apostrophe_s_contraction` (no short form),
- `--preferred_name_sources` (short form `-W`), and
- `--truncate_candidates_mappings` (short form `-X`).