MetaMap2016 Usage Notes

François-Michel Lang
metamap@nlm.nih.gov

January 31, 2016

This document explains MetaMap's command-line options, which support a wide variety of processing. All options have a long name (e.g., `--term_processing`), and most have a short name (e.g., `-z`) as well, for simplicity and ease of use.

All use of MetaMap requires a UMLS Metathesaurus license; see this page for all access to MetaMap, including interactive and batch use from our website and downloading and running it locally at user sites.

The MetaMap 2016 Release Notes are available here. Users are encouraged to review the MetaMap Usage FAQ, which presents many use cases and scenarios, here.

Click on any of the following links for documentation about the various types of MetaMap options.

Usage

There are two ways to use MetaMap interactively, reading input text from the keyboard and seeing output on the screen:

1. `metamap [ options ]`
   then type your input text, e.g., `lung cancer`, at the "`|:`" prompt.
2. `echo lung cancer | metamap [ options ]`

For processing an input file:

```
metamap [ options ] InputFile OutputFile
```

The `InputFile` and `OutputFile` options, if specified, must be the last two arguments. If `OutputFile` is not specified, it will default to `InputFile.out`. Note that if the output file (whether specified on the command line or not) already exists, it will be overwritten and its original contents lost.

For processing another program's output:

```
OtherProgram | metamap [ options ]
OtherProgram | metamap [ options ] > OutputFile
```

To generate a short list of all MetaMap options, simply call

```
metamap --help
```

Data Options

MetaMap's data options determine the UMLS Metathesaurus release, the data model, and subset of UMLS source vocabularies to use.

`-Z (--mm_data_year)` `<release>`
Sets the version of the UMLS Metathesaurus to use, e.g., 2015AA, 2015AB, etc.

`-A (--strict_model)` (default)
`-C (--relaxed_model)`
Sets MetaMap's data model (strict or relaxed). See this page for more information about MetaMap's strict and relaxed data models.

`-V (--mm_data_version)` `<version>`
Sets MetaMap's data version (Base, USAbase (the default), and NLM). See this page for more information about MetaMap's Base, USAbase, and NLM data versions.

Processing Options

Processing options control MetaMap's search algorithms and therefore affect the choice of UMLS concepts identified.

`--UDA` `<file>`
Allows users to specify acronyms and abbreviations (AAs) that are not defined in the input text ("UDA" is a recursive acronym meaning "user-defined AA"). This option is designed specifically for processing clinical text, which often contains undefined AAs. See this page for more information about processing clinical text with MetaMap. More information about MetaMap's UDA processing is available here.

`--blanklines` `<integer>`
MetaMap by default treats any number of consecutive blank lines as an input-record separator. However, this default behavior is often undesirable for analyzing clinical text, because clinical notes often include multiple blank or whitespace lines in the middle of reports. See this page for more information about processing clinical text with MetaMap.

`--cascade` New in MetaMap2016
Causes MetaMap to ignore concepts that overlap textually with any candidate that is excluded because of Semantic Types or Sources (cascaded deletions). An explanatory example is necessary.

MetaMap by default maps the input text *logistic regression* to the UMLS concept `Logistic Regression`, whose Semantic Type (ST) is Research Activity (short form `resa`). Another UMLS concept identified is `Regression`, whose ST is Disease or Syndrome (`dsyn`), but that concept does not appear in MetaMap's final mappings, because `Logistic Regression` receives a higher score.

Suppose now we restrict MetaMap to concepts whose ST is Disease or Syndrome (`dsyn`) by using the option `-J dsyn`, described below. The concept `logistic regression` that would otherwise be found will then be excluded because its ST is Research Activity (`resa`). MetaMap's final mapping will then contain the smaller concept `Regression`, because it is a `dsyn`, and therefore not excluded.

Using the `--cascade` option will also then rule out the concept `Regression` even though it is a `dsyn` because it overlaps with the excluded concept `Logistic Regression`.

`--negex_st_add` `<list>`
`--negex_st_del` `<list>`
`--negex_st_set` `<list>`
`--negex_trigger_file` `<file>`
Allows customization of UMLS Semantic Types used in NegEx processing. See this page for more information about the first three options and Section 7 on this page for more information about the last option.

`--nomap` `<file>` New in MetaMap2016
Causes MetaMap to block user-specified infelicitous mappings. See this page for more detailed information.

`--sldi`
`--sldiID`
Causes MetaMap to read in a list of terms (`--sldi`) or a list of terms with IDs (`--sldiID`) rather than free text.

`--prune` `<integer>`
Specify the maximum number of candidate concepts used in creating final mappings. This option should be used only if MetaMap runs for a very long time.

`-@ (--WSD)` `<hostname>`
Specify the hostname running the WSD Server to be used for word-sense disambiguation.

`-a (--all_acros_abbrs)`
Allows the use of any acronym/abbreviation (AA) variants, which are the least reliable form of variation because of the extreme ambiguity of AA variants.

`-b (--compute_all_mappings)`
Causes MetaMap to compute and display all mappings, rather than only the top-scoring ones. Note: It is rarely useful to display all mappings because of their large number.

`-d (--no_derivational_variants)`
Prevents the use of any derivational variation in the computation of word variants. This option exists because derivational variants can involve a significant change in meaning.

`-D (all_derivational_variants)`
Allow the use of *all* derivational variation, instead of only those between adjectives and nouns (the default). Adjective/noun derivational variants are generally the best derivational variants.

`-e (--exclude_sources)` `<list>`
Excludes the UMLS Source Vocabularies specified in the comma-separated list while mapping concepts. E.g., `-e ICD10CM,MSH`. More information about UMLS Source Vocabularies is available here.

`-g (--allow_concept_gaps)`
Causes MetaMap to retrieve Metathesaurus candidates with gaps. For example, with this option, MetaMap maps the text *chronic toxicity* to the UMLS concept *chronic radiation toxicity*. The word *radiation* is inserted into the gap between *chronic* and *toxicity*. This option does not appreciably affect MetaMap's performance, and is best suited for browsing purposes.

`-i (--ignore_word_order)`
Allows MetaMap to ignore the order of words in the input text. MetaMap was originally developed to process full text, and consequently depended very strongly on normal English word order. This option avoids the use of specialized word indexes used for efficient candidate retrieval; it also ignores word order when matching phrase text to candidate words; and it replaces the normal coverage metric with an involvement metric for evaluating how well a candidate covers the words of a phrase. Using this option tends to increase recall but decrease precision.

`-J (--restrict_to_sts)` `<list>`
Restricts output to those concepts with one of the semantic types specified in the comma-separated list. E.g., `-J dsyn,neop`. More information about UMLS Semantic Types is available here.

`-k (--exclude_sts)` `<list>`
Excludes concepts not having a semantic type in the comma-separated list. E.g., `-k dsyn,neop`. More information about UMLS Semantic Types is available here.

`-l (--allow_large_n)`
Enables retrieval of Metathesaurus candidates for (a) two-character words occurring in more than 4,000 Metathesaurus strings and (b) one-character words occurring in more than 2,000 Metathesaurus strings. This option also allows retrieval for words that can be a preposition, conjunction or determiner.

`-L (--lexicon_year)` `<year>`
Specifies which version of the SPECIALIST lexicon to use. MetaMap defaults to the most recent lexicon, which is associated with the most recent UMLS release. If this default option is overridden (e.g., `-L 2013`), we recommend overriding the default UMLS release as well with a UMLS version of the same year (e.g., `-Z 2013AB`).

`-o (--allow_overmatches)`
Causes MetaMap to retrieve Metathesaurus candidates containing words on one or both ends that do not match the text. For example, overmatches of *medicine* include *Antibiotic Medicine*, *Medicine Preparations*, *Investigational Medicinal Product*. This option *greatly* increases the number of candidates retrieved and is consequently much slower than MetaMap without overmatches. It is most appropriate for MetaMap's Browse Mode.

`-Q (--composite_phrases)` `<integer>`
Causes MetaMap to construct longer, composite phrases from the smaller phrases produced by the parser; the integer operand specifies the number of prepositional phrases that can be glommed onto the initial noun phrase. This option is on by default with a setting of 4, but can be overridden (e.g., `-Q 2` or `-Q 0`) to achieve greater processing efficiency, albeit possibly with less good results. For more information, see this page.

`-r (--threshold)` `<integer>`
Restricts output to UMLS candidate concepts whose evaluation score equals or exceeds the specified threshold. Judicious use of this option can exclude false positives when some input text has no close matches in the Metathesaurus. An appropriate threshold can usually be determined simply by examining MetaMap output for typical text in a given application.

`-R (--restrict_to_sources)` `<list>`
Uses only the specified UMLS Source Vocabularies while mapping concepts. E.g., `-R ICD10CM,MSH`. More information about UMLS Source Vocabularies is available here.

`-S (--TAGGER_SERVER)` `<hostname>`
Specifies the hostname running the Tagger Server to be used for part-of-speech tagging. See the very next option for more information.

`-t (--no_tagging)`
Bypasses the part-of-speech tagger server. By default, the SPECIALIST parser will use the results of a tagger to assist in parsing. MetaMap currently uses the MedPost/SKR tagger. See this page and this page for more information about the MedPost tagger, which was developed at NCBI specifically for tagging biomedical text; we modified it to use our part-of-speech tags.

`-u (--unique_acros_abbrs_only)`
Restricts the generation of acronym/abbreviation (AA) variants to those forms with unique expansions. This option generally produces better results than allowing all forms of acronym/abbreviation variants (using `-a` or `all_acros_abbrs`), but our experience has shown that still better results are produced by allowing *no* AA variants.

`-y (--word_sense_disambiguation)`
Causes MetaMap to attempt to disambiguate among concepts scoring equally well. More information about MetaMap's WSD is available here.

`-Y (--prefer_multiple_concepts)`
Causes MetaMap to score mappings with more concepts higher than those with fewer concepts (simply by inverting the normal cohesiveness value). For example, with this option, the input text *lung cancer* will be mapped to the two concepts `Lung` and `Cancer`, rather than the single concept `Lung Cancer`. This option is useful for discovering semantic relationships among concepts found in text (e.g., `lung-LOCATION_OF-cancer`).

`-z (--term_processing)`
Process terms, i.e., short text fragments, rather than a document containing complete sentences. See this page for more information about term processing. A typical use of term processing involves processing a list of terms or a list of terms with ID.

Output Formats

MetaMap's default output format is human-readable output, which is best suited for learning about MetaMap and testing various strategies. Other output formats such as Prolog Machine Output, XML Output, and Fielded MMI Output are far better suited for automated downstream postprocessing.

`-q` (`--machine_output`)
Generates Prolog terms rather than human-readable form. See this page for more information about MetaMap's Prolog Machine Output.

`--XMLf`
Generates formatted XML, one XML document per input record/citation. Formatted XML is suitable for reading by humans, but more space intensive than unformatted XML. See this page for detailed information about MetaMap's XML output formats.

`--XMLn`
Generates unformatted XML, one XML document per input record/citation. Formatted XML is not suitable for reading by humans, but more compact than formatted XML. See this page for detailed information about MetaMap's XML output formats.

`--XMLf1`
Generates formatted XML, one XML document per input file. See this page for detailed information about MetaMap's XML output formats.

`--XMLn1`
Generates unformatted XML, one XML document per input file. See this page for detailed information about MetaMap's XML output formats.

`-N` (`--fielded_mmi_output`)
Generate Fielded MMI (MetaMap Indexing) output. See this page for detailed information about MetaMap's MMI output.

Output Options

MetaMap provides a wide variety of options that control its output. The options that affect only MetaMap's human-readable output; are labeled "HR only"; using those options with any output format other than human-readable will generate a warning, or, in certain cases, an error.

`--aas_only`
Limits MetaMap's output to information about acronyms and abbreviations (AAs); using this option does no named-entity recognition, and is therefore much faster than regular MetaMap; HR only.

`-+` (`--bracketed_output`)
Surrounds the Phrase, Candidates, and Mappings section of output with >>>>> and <<<<< brackets; HR only. E.g.,

```
>>>>> Phrase
heart attack
<<<<< Phrase
```

`-c` (`--show_candidates`)
By default, MetaMap output contains only final mappings, and not all the candidate concepts found in the text. This option forces the display of all Metathesaurus candidate concepts identified in the text, regardless of whether they appear in MetaMap's final mappings. Candidates are displayed best to worst, according to the MetaMap evaluation metric.

`-E` (`--indicate_citation_end`)
Causes the end-of-transmission term `'EOT'` to be generated (followed by a period) at the end of the output stream. This option is useful for processing using our Batch Scheduler with validated processing, or for ensuring that MetaMap did in fact complete processing an input file.

`-f` (`--number_the_mappings`)
Numbers the final mappings; HR only.

`-F` (`--formal_tagger_output`)
Displays the tagging information returned by the tagger server.

`-G` (`--sources`)
Displays the Metathesaurus sources for each candidate and mapping in the output; HR only. More information about UMLS Source Vocabularies is available here.

`-I` (`--show_cuis`)
Displays the UMLS CUI for each concept; HR only.

`-j` (`--dump_aas`)
Displays the acronyms/abbreviations (AAs) discovered by MetaMap in the form below (pretty-printed for readability); HR only.

```
AA | PMID | Acronym | Expansion | #Acronym Tokens | #Acronym Chars |
             #ExpansionTokens | #Expansion Chars | Text Offsets
```

E.g., for the input *confidence interval (CI)*, MetaMap would display

```
AA|00000000|CI|confidence interval|1|2|3|19|21:2
```

`-m` (`--hide_mappings`)
By default, MetaMap output contains only final mappings, and not all the candidate concepts found in the text. This option disables the display of mappings. It is an error to use this option without `-c --show_candidates`).

`-n` (`--number_the_candidates`)
Numbers the candidates in a displayed candidate list; HR only.

`-p` (`--hide_plain_syntax`)
Disables the display of the words forming each phrase, as determined by the SPECIALIST parser; HR only.

`-s (--short_semantic_types)`
Displays the short form of UMLS Semantic Types rather than the long form, e.g., `dsyn` rather than `Disease or Syndrome`; HR only.

`-T (--tagger_output)`
Displays the output of the MedPost/SKR tagger lining up input words on one line with their tags on a line below.

`-v (--variants)`
Displays the variants generated for each input word.

`-x (--syntax)`
Displays the output of the SPECIALIST parser; HR only.

`--negex`
Displays information about negated UMLS concepts occurring in the input and the associated strings that caused the negation; HR only. Negation information is always included in Prolog Machine Output, XML Output, and Fielded MMI Output.

`--silent`
Suppresses the display of header information such as that shown below.

```
Berkeley DB databases (USAbase 2015AB strict model) are open.
Static variants will come from table varsan in
       /nfsvol/nls/II_Group_WorkArea/MetaMap_DB//DB.USAbase.2015AB.strict.
Derivational Variants: Adj/noun ONLY.
Variant generation mode: static.
Established connection $stream(140152552284000) to TAGGER Server on ii-server3.

a.out.Linux (2015)

Control options:
  composite_phrases=4
  lexicon=db
  mm_data_year=2015AB
```