Using Additional Data Sets with Public MetaMap

Willie Rogers

January 13, 2023

1 Archives files necessary for using an additional data set

For each supported UMLS release, MetaMap provides three data subsets (Base, USAbase, and NLM), each subset is derived from a subset of the UMLS Metathesaurus; these subsets are explained in Section 3 of MetaMap 2011 Release Notes (http://metamap.nlm.nih.gov/Docs/MM_2011_ReleaseNotes.pdf). Each of these three data subsets are available in a both strict and relaxed models see Note on Model Documentation (§ 4) at the end of this page.

To use any model for a particular release you need two archive files: The base archive for the release and the archive file for the particular model you want to use . If you wish to use more than one model you only need to download the base archive file once.

If you wish to use the relaxed model for Base data set 2012AA it is necessary to download the base archive file: `public_mm_data_base_2012aa_base.bz2` and the relaxed model archive: `public_mm_data_base_2012aa_relaxed.bz2`. You will also need the file `public_mm_data_dblexicon_2012.tar.bz2` that provides the 2012 version of the SPECIALIST Lexicon required for the 2012aa data set.

2 Installing the data sets for additional models

After downloading the necessary archive files for additional data set, first move to the directory containing the directory of the existing public_mm installation and then extract the additional data set using bzip2 and tar.

For example, to add the 2012AA Base relaxed model to your MetaMap installation:

```
$ cd <directory containing existing public_mm installation>
$ tar xvfj public_mm_data_dblexicon_2012.tar.bz2
$ tar xvfj public_mm_data_base_2012aa_base.tar.bz2
$ tar xvfj public_mm_data_base_2012aa_relaxed.tar.bz2
```

The 2012aa data-files should be installed now, to use the new data set run MetaMap with the options `-Z <4 digit year and 2 letter release> -V <UMLS subset> -<model name>_model` in the current case the options `-Z 2012AA -V Base -relaxed_model` will suffice:

```
$ cd public_mm
$ echo "lung cancer" | ./bin/metamap13 -Z 2012AA -V Base --relaxed_model
```

The output should be similar to this:

```
...
metamap13.BINARY.Linux (2013)

Control options:
  composite_phrases=4
  lexicon=db
  mm_data_year=2012AA
  mm_data_version=Base
  relaxed_model
Processing 00000000.tx.1: lung cancer

Phrase: "lung cancer"
Meta Candidates (Total=10; Excluded=3; Pruned=0; Remaining=7)
  1000   Lung Cancer (Malignant neoplasm of lung) [Neoplastic Process]
  1000   LUNG CANCER (Carcinoma of lung) [Neoplastic Process]
   861   Cancer (Malignant Neoplasms) [Neoplastic Process]
   861   Lung [Body Part, Organ, or Organ Component]
   861   Lung (Lung diseases) [Disease or Syndrome]
   861   Cancer (Cancer Genus) [Eukaryote]
   861   Cancer (Specialty Type - cancer) [Biomedical Occupation or Discipline]
   805 E Pulmonary (Pulmonary (qualifier value)) [Qualitative Concept]
   768 E Pneumonia [Disease or Syndrome]
   768 E Pulmonary Arteries (Pulmonary artery structure) [Body Part, Organ, or Organ Component]
Meta Mapping (1000):
  1000   LUNG CANCER (Carcinoma of lung) [Neoplastic Process]
Meta Mapping (1000):
  1000   Lung Cancer (Malignant neoplasm of lung) [Neoplastic Process]
$
```

3 Archive Filename Organization

The archive files are named using the following nomenclature:
public_mm_data_{**UMLS subset**}_{**UMLS year**}{**UMLS release**}_{**model**}.tar.bz2
On Windows:
public_mm_data_win32_{**UMLS subset**}_{**UMLS year**}{**UMLS release**}_{**model**}.7z
public_mm_data_win32_{**UMLS subset**}_{**UMLS year**}{**UMLS release**}_{**model**}.zip
A description of the segments of the archive name follows:

UMLS subset Subset of UMLS Meta-Thesaurus (Base, USAbase, and NLM)

UMLS year The year the UMLS Meta-Thesaurus release was created.

UMLS release AA - first release, AB - second release

Below are the MetaMap UMLS subsets and the corresponding UMLS Metathesaurus subsets used to generate them:

Base Level 0 vocabularies only.

USAbase Level 0 vocabularies + SNOMEDCT

NLM All vocabularies except (CDT, CPT, and other HCPCS vocabularies)

See MetamorphoSys Help for additional information: http://www.nlm.nih.gov/research/umls/implementation_resources/metamorphosys/help.html

4 Note: Model Documentation

For a description of the content of each model, see the paper: "Filtering the UMLS Metathesaurus for MetaMap" at the SKR website under "Research Information" (http://ii.nlm.nih.gov/Publications/index.shtml).