Using Additional Datasets with Public MetaMap

Willie Rogers

August 30, 2013

1 Archives files necessary for using an additional dataset

For each supported UMLS release, MetaMap provides three data versions (Base, USAbase, and NLM), which are explained in Section 3 of MetaMap 2011 Release Notes (`http://metamap.nlm.nih.gov/MM_2011_ReleaseNotes.pdf`). Each of these three versions is available in a both strict and relaxed versions see Note on Model Documentation (§ 3) 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 2006 it is necessary to download the base archive file: `public_mm_data_2006_base.bz2` and the relaxed model archive: `public_mm_data_2006_relaxed.bz2`.

2 Installing the datasets for additional models

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

For example, to add the 2006 relaxed model to your MetaMap installation:

```
$ cd <directory containing existing public_mm installation>
$ bzip2 -dc public_mm_data_2006_base.tar.bz2 | tar xf -
$ bzip2 -dc public_mm_data_2006_relaxed.tar.bz2 | tar xf -
```

The 2006 datafiles should be installed now, to use the new dataset run metamap with the options `-Z <two digit year> -<model name>_model` in the current case the options `-Z 06 -relaxed_model` will suffice:

```
$ cd public_mm
$ echo "lung cancer" | ./bin/metamap08 -Z 06 --relaxed_model
```

The output should be similar to this:

```
MetaMap (2008)

Control options:
  mm_data_year=06
  relaxed_model
Berkeley DB databases (normal relaxed 06 model) are open.
Static variants will come from table varsan.
Accessing lexicon <parent directory>/public_mm/lexicon/data/lexiconStatic2008.
Variant generation mode: static.
Initializing tagger on localhost...

Processing 00000000.tx.1: lung cancer

Phrase: "lung cancer"
Meta Candidates (10):
  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 LUNG (Lung Problem) [Disease or Syndrome]
   861 Cancer (Cancer Genus) [Invertebrate]
   861 Lung (Entire lung) [Body Part, Organ, or Organ Component]
   861 Cancer (Specialty Type - cancer) [Biomedical Occupation or Discipline]
   768 Pneumonia [Disease or Syndrome]
Meta Mapping (1000):
  1000 Lung Cancer (Carcinoma of lung) [Neoplastic Process]
Meta Mapping (1000):
  1000 Lung Cancer (Malignant neoplasm of lung) [Neoplastic Process]
$
```

3 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://skr.nlm.nih.gov/papers/index.shtml) .