MetaMap2020 Release Notes

François-Michel Lang
metamap@nlm.nih.gov
May 20, 2020

MetaMap20 includes a number of significant enhancements, described in the rest of this document. These enhancements are currently available only with the downloaded and locally installed version of MetaMap20; they will eventually be available on the version of MetaMap running on our websites, e.g., here and here.

1 UTF-8

MetaMap has historically required ASCII input because its UMLS data was always pure ASCII. However, we included UTF-8 strings in MetaMap's 2020AA UMLS data, and MetaMap20 will accordingly accept UTF-8 input. See the Unformatted English Text FAQ for additional information. We have also added ASCII equivalents to all non-ASCII UMLS strings to potentially improve recall. For example, in the 2020AA MRCONSO.RRF file, the following CUIs (among many others) contain exclusively UTF-8 English strings:

```
C1398762|Glénard
C1505197|Ricridène
C1530290|Hämatopan
C3825987|Mâeniáere's disease
```

We therefore added the following ASCII strings to the 2020AA MetaMap data:

```
C1398762|Glenard
C1505197|Ricridene
C1530290|Hamatopan
C3825987|Maeniaere's disease
```

This addition will allow MetaMap to map both *glénard* and *glenard* to CUI C1398762.

The release of MetaMap20 also includes an update to the MedPost/SKR Part-of-Speech Tagger, which was also modified to handle UTF-8 text.

We strongly encourage the use of MetaMap20 *only* with this updated tagger, and, similarly, use of the updated tagger *only* with MetaMap20.

2 Graceful Degradation

All previous versions of MetaMap had thrown a fatal error and aborted on those rare occasions when it couldn't process some text in an input file. We have added a graceful-degradation mechanism that handles certain formerly-fatal errors by instead printing a warning to stderr, skipping the rest

of the problematic utterance, and silently proceeding to the next utterance. No output will be generated for the remainder of the utterance, but on the other hand MetaMap won't abort either. This change will be especially useful for a MetaMap invocation to process an input file containing multiple citations.

3 Small Change to User-Defined Acronyms/Abbreviations

User-defined acronyms/abbreviations (UDAs) were introduced in Section 9 of the MetaMap 2011 Release Notes and also described in the Clinical Text FAQ.

Formerly, in a line in the UDA file like "`x|y`", the shorter of the two strings was treated as the acronym/abbreviation (AA), and the longer, as the expansion or replacement—regardless of their order in the line.

Beginning with MetaMap20, the first string is taken as the AA, and the second, the expansion or replacement—regardless of the relative lengths of the two strings. This change allows users to tell MetaMap to replace any string with any other string independently of their relative lengths.

Thia "off-label" use of the UDA mechanism is similar to the strategy described in Blocking Unwanted UMLS Concepts FAQ.