i2b2 Medication Extraction Challenge

Input/Output Formats

As of 6/10

(Subject to change, with notice)

Ozlem Uzuner, Imre Solti, and Fei Xia

The input to the systems will be discharge summaries. These summaries are completely de-identified and xmlized in order to separate the records from each other. The input files contain no other markups.

For each file, the output to be created will be a list of medications. For each listed medication, the following information needs to be included:

1. medication name and its offset (marker “m”)
2. dosage and its offset (marker “do”)
3. mode/route of administration and its offset (marker “mo”)
4. frequency and its offset (marker “f”)
5. duration and its offset (marker “du”)
6. reason and its offset (marker “r”)
7. event (marker “e”)
8. temporal marker (marker “t”)
9. certainty (marker “c”)
10. found in list/narrative of the text (marker “ln”)

Each entry should consist of the above 10 fields. Each entry should be printed on its own line. The fields are separated by the “||” (double pipe) character:

Medication name & offset||dosage & offset||mode & offset||frequency & offset||duration & offset||reason & offset||event|| temporal marker||certainty||list/narrative

Each field follows the format:

…||m=“medication name” line number:start token position of medication name line number:end token position of medication name ||…

Line numbers start from 1 on the first line of the file and indicate the line number at which the field is found. Token positions start with 0 at the beginning of each line and mark the first token that is the field. Tokens are determined by white space.

For fields that are not mentioned in the text, the field should be set to “nm” and there should be no offset.

The output of each system should be returned to i2b2 in a single zip file. The zip file should contain an individual output file of annotations for each input data file. The name of the output file for each input should contain the name of the input file. E.g., for input file 11995, the output file should be named 11995.output. The list of entries for each record can be unordered within itself. However, the lists for different records cannot be mixed, merged, or otherwise combined.

Sample output file, e.g., file 42341231.output:

```
m=“oxygen” 5:1 5:1||do=“nm”||mo=“nm”||f=“nm”||du=“nm”||
r=“respiratory distress” 5:10 5:11|| …
m=“nitroglycerin” 5:2 5:2||do=“nm”||mo=“nm”||f=“nm”||du=“nm”||
r=“respiratory distress” 5:10 5:11|| …
m=“aspirin” 5:3 5:3||do=“nm”||mo=“nm”||f=“nm”||du=“nm”||
r=“respiratory distress” 5:10 5:11|| …
m=“lasix” 5:4 5:4||do=“nm”||mo=“nm”||f=“nm”||du=“nm”||
r=“respiratory distress” 5:10 5:11|| …
m=“cpap” 5:5 5:5||do=“nm”||mo=“nm”||f=“nm”||du=“nm”|| r=“respiratory
distress” 5:10 5:11|| …
m=“morphine” 5:6 5:6||do=“nm”||mo=“nm”||f=“nm”||du=“nm”||
r=“respiratory distress” 5:10 5:11|| …
```

Sample output file, e.g., file 3121.output:

```
m=“heparin” 7:11 7:11||do=“50 mg” 7:12 7:13||mo=“intravenous” 7:14
7:14||f=“nm”||du=“nm”||r=“nm”||e=“stop”|| …
m=“heparin” 10:17 10:17||do=“25 mg” 10:18 10:19||mo=“intravenous”
10:20 10:20||f=“nm”||du=“nm”||r=“nm”||e=“start”|| …
```