**Compiling chemicals/drugs lexicon from UMLS Metathesaurus, DrugBank and PharmGKB**

**UMLS Metathesaurus (2015AB version)**

UMLSChemicalsDrugsExtractor.java is meant to extract the chemicals/drugs concepts and their synonyms from the processed UMLS Metathesaurus (refer to our project ConceptMap:UMLSMetathesaurusCompiler). The program can be executed with the following command:

*$ javac UMLSChemicalsDrugsExtractor.java*

*$ java UMLSChemicalsDrugsExtractor INPUT\_FILE OUTPUT\_FILE\_umls*

UMLSSynonymsCollector.java groups chemicals/drugs and their synonyms into individual records (or rows). This file was used to combine chemicals/drugs resources from DrugBank and PharmGKB.

*$ javac UMLSSynonymsCollector.java*

*$ java UMLSSynonymsCollector OUTPUT\_FILE\_umls OUTPUT\_FILE\_umls\_grouped*

**DrugBank**

DrugNameSynonymExtractor.java is meant to extract the drug names and synonyms from the XML file downloaded from DrugBank. The program can be executed with the following command:

*$ javac DrugNameSynonymExtractor.java*

*$ java DrugNameSynonymExtractor INPUT\_FILE\_xml OUTPUT\_FILE1\_ drugbankID\_drugs\_synonyms OUTPUT\_FILE2\_ salts\_info OUTPUT\_FILE3\_ drugbankID\_drugsalts*

While, UMLS Metathesaurus contains semantic type (TUI) for the concepts, DrugBank and PharmGKB does not contain TUI. We defined TUI for DrugBank entities based on the groups assigned to each drug. DrugBank includes the following groups: approved, experimental, investigational, illicit, nutraceutical and withdrawn. Table 1 lists TUIs assigned to each set of groups identified in DrugBank.

Table 1. Drug groups and TUI

|  |  |
| --- | --- |
| **Drug Group** | **TUI** |
| Approved (and related) | T220 to T231 |
| Experimental (and related) | T330 to T333 |
| Investigational (and related) | T440 to T444 |
| Ilicit (and related) | T550 to T551 |
| Neutraceutical | T660 |
| Withdrawn | T000 |
| Drug Salts | T999 |

Each drug is annotated with one or more drug groups. All combinations of drug groups can be retrieved using DrugBankGroupExtractor.java. We defined TUI for each combination (Table 2).

*$ javac DrugGroupExtractor.java*

*$ java DrugGroupExtractor INPUT\_FILE\_xml OUTPUT\_FILE\_drugbankID\_group*

Table 2: Combination of drug groups and TUI

|  |  |  |
| --- | --- | --- |
| **Drug Group (combinations)** | **Number of Drugs** | **TUI** |
| approved | 1,590 | T220 |
| approved, experimental | 2 | T221 |
| approved, illicit | 55 | T222 |
| approved, illicit, investigational | 6 | T223 |
| approved, illicit, investigational, withdrawn | 1 | T224 |
| approved, illicit, withdrawn | 8 | T225 |
| approved, investigational | 377 | T226 |
| approved, investigational, nutraceutical | 3 | T227 |
| approved, investigational, withdrawn | 12 | T228 |
| approved, nutraceutical | 67 | T229 |
| approved, nutraceutical, withdrawn | 1 | T230 |
| approved, withdrawn | 71 | T231 |
| experimental | 4,919 | T330 |
| experimental, illicit | 92 | T331 |
| experimental, illicit, withdrawn | 1 | T332 |
| experimental, investigational | 10 | T333 |
| illicit, investigational | 1 | T440 |
| illicit, investigational, withdrawn | 1 | T441 |
| investigational | 833 | T442 |
| investigational, nutraceutical | 1 | T443 |
| investigational, withdrawn | 8 | T444 |
| illicit | 26 | T550 |
| illicit, withdrawn | 7 | T551 |
| nutraceutical | 21 | T660 |
| withdrawn | 90 | T000 |
| drug salts\*\* | 1,201 | T999 |

\*\* Salts in DrugBank does not contain group information. Yet, we define TUI.

DrugBankGroupTUIMatcher.java assigns TUI to each Drug ID. The following commands are used to execute the program.

*$ javac DrugBankGroupTUIMatcher.java*

*$ java DrugBankGroupTUIMatcher OUTPUT\_FILE\_drugbankID\_group OUTPUT\_FILE\_drugbankID\_TUI*

DrugBankTUIAssigner.java replaces DrugBank drug\_ID with drug\_ID\_TUI and salt\_ID with salt\_ID\_TUI.

*$ javac DrugBankTUIAssigner.java*

*$ java DrugBankTUIAssigner OUTPUT\_FILE\_drugbankID\_TUI OUTPUT\_FILE1\_ drugbankID\_drugs\_synonyms OUTPUT\_FILE\_drugbankID\_TUI\_drugsAndSalts\_synonyms*

**PharmGKB**

The download drugs.zip file from PharmGKB is a tab-separated file. It contains 10 columns of information:

PharmGKB\_Accession\_Id | Name | Generic\_Name | Trade\_Name | Brand\_Mixtures | Type | Cross-reference | SMILES | Dosing\_Guideline | External\_Vocabulary.

To create the dictionary, we consider the following 5 columns:

PharmGKB\_Accession\_Id | Name | Generic\_Name | Trade\_Name | Cross-reference.

*$ cut -f1,2,3,4,7 drugs\_unzipped > OUTPUT\_FILE\_PharmGKB\_drugs\_synonyms*

DrugSynonymExtractor.java is developed to process the generic names and trade names, and to extract PharmGKB\_Accession\_ID. The generic and trade names are separated by three ashes (i.e. "###"). The format of the output file is PharmGKB\_Accession\_ID\tName\tGeneric\_names\tTrade\_names\tCross\_references. The following command is used for executing the program.

*$ javac DrugSynonymExtractor.java*

*$ java DrugSynonymExtractor OUTPUT\_FILE\_PharmGKB\_drugs\_synonyms OUTPUT\_FILE\_pharmgkb*

We define drugs from PharmGKB with a single TUI, T777. PharmGKBTUIAssigner.java assigns TUI to each drug. The following commands are used to execute the program.

*$ javac PharmGKBTUIAssigner.java*

*$ java PharmGKBTUIAssigner OUTPUT\_FILE\_PharmGKB\_drugs\_synonyms OUTPUT\_FILE\_PharmGKB\_TUI\_drugs\_synonyms*

**Dictionary compiling**

**Synonyms**

Our approach considers UMLS Metathesaurus as the main resource and expands the synonyms using DrugBank and PharmGKB. The dictionary compilation thus requires list of synonyms from DrugBank and PharmGKB. DrugBankSynonymsList.java collects synonyms from DrugBank and PharmGKBSynonymsList.java collects synonyms from PharmGKB. The following commands are used to execute the programs.

*$ javac DrugBankSynonymsList.java*

*$ java DrugBankSynonymsList OUTPUT\_FILE1\_ drugbankID\_drugs\_synonyms OUTPUT\_FILE\_drugbankSynonyms*

*$ javac PharmGKBSynonymsList.java*

*$ java PharmGKBSynonymsList OUTPUT\_FILE\_ pharmgkb\_drugs\_synonyms\_output OUTPUT\_FILE\_pharmgkbSynonyms*

**Normalization of drug names**

Drug mentions across the resources may not be same always. Therefore, it is mandatory to normalize the drug names.

UMLSDrugsNormalizer.java normalizes the chemicals and drugs in UMLS Metathesaurus. The following commands are used to execute the program.

*$ javac UMLSDrugsNormalizer.java*

*$ java UMLSDrugsNormalizer OUTPUT\_FILE\_umls OUTPUT\_FILE\_umlsSynonyms\_normalized*

DrugBankDrugsNormalizer.java normalizes the drugs in DrugBank. The following commands are used to execute the program.

*$ javac DrugBankDrugsNormalizer.java*

*$ java DrugBankDrugsNormalizer OUTPUT\_FILE\_drugbankSynonyms OUTPUT\_FILE\_drugbankSynonyms\_normalized*

PharmGKBDrugsNormalizer.java normalizes the drugs in PharmGKB. The following commands are used to execute the program.

*$ javac PharmGKBDrugsNormalizer.java*

*$ java PharmGKBDrugsNormalizer OUTPUT\_FILE\_pharmgkbSynonyms OUTPUT\_FILE\_pharmgkbSynonyms\_normalized*

**Compiling loops**

Loop 1 – UMLS > DrugBank > PharmGKB

Our lexicon compiling is carried out in three stages: First, the chemical/drug concepts and synonyms are collected from UMLS Metathesaurus. For each concept, the program scans DrugBank and PharmGKB for synonyms that are not in UMLS Metathesaus, and include them as synonyms. DrugsDictionaryCompilerLoop1.java is the related Java program file and can be executed with the following command:

*$ javac DrugsDictionaryCompilerLoop1.java*

*$ java DrugsDictionaryCompilerLoop1 OUTPUT\_FILE\_drugBankSynonyms OUTPUT\_FILE\_pharmgkbSynonyms OUTPUT\_FILE\_drugbankDrugs\_normalized OUTPUT\_FILE\_pharmgkbDrugs\_normalized OUTPUT\_FILE\_umls\_grouped OUTPUT\_FILE1\_drugsDictionaryLoop1 OUTPUT\_FILE2\_synonymsCountLoop1*

A list of DrugBank drugs (i.e. DrugBank drug ID) mapping to chemicals/drugs from UMLS Metathesaurus is obtained with the following Linux command:

*$ cut -f6 OUTPUT\_FILE1\_drugsDictionaryLoop1.txt | sort | uniq > UMLS2DrugBankID\_List*

A list of PharmGKB drugs (i.e. PharmGKB drug ID) mapping to chemicals/drugs from UMLS Metathesaurus is obtained with the following Linux command:

*$ cut -f7 OUTPUT\_FILE1\_drugsDictionaryLoop1.txt | sort | uniq > UMLS2PharmGKBID\_List*

Loop 2 – DrugBank > PharmGKB

DrugBank Drugs that were not mapped to UMLS Chemicals & Drugs are collected using DrugBankID4Loop2.java and used as input for Loop2. The following commands are used to execute the program.

*$ javac DrugBankID4Loop2.java*

*$ java DrugBankID4Loop2 UMLS2DrugBankID\_list OUTPUT\_FILE\_drugbankID\_TUI\_drugsAndSalts\_synonyms OUTPUT\_FILE\_drugbankDrugsForloop2*

The drug concepts that are not concepts or synonyms of UMLS Metathesaurus are filtered from DrugBank. For each concept, the program collects the synonyms from DrugBank and scans PharmGKB for synonyms that are not in DrugBank. DrugsDictionaryCompilerLoop2.java is the related Java program file and can be executed with the following command:

*$ javac DrugsDictionaryCompilerLoop2.java*

*$ java DrugsDictionaryCompilerLoop2 OUTPUT\_FILE\_pharmgkbSynonyms OUTPUT\_FILE\_pharmgkbDrugs\_normalized OUTPUT\_FILE\_drugbankDrugsForloop2 OUTPUT\_FILE1\_drugsDictionaryLoop2 OUTPUT\_FILE2\_synonymsCountLoop2*

A list of PharmGKB drugs (i.e. PharmGKB drug ID) mapping to drugs from DrugBank is obtained with the following Linux command:

*$ cut -f7 OUTPUT\_FILE1\_drugsDictionaryLoop2.txt | sort | uniq > DrugBank2PharmGKBID\_List*

Loop 3 - PharmGKB

A list of PharmGKB drugs mapping to UMLS Metathesaurus and DrugBank is obtained with the following Linux command:

*$ cat UMLS2PharmGKBID\_List DrugBank2PharmGKBID\_List > PharmGKB\_drugs\_loop1\_loop2*

PharmGKB Drugs that were not mapped to UMLS Metathesaurus and DrubBank are collected using PharmGKBDrugs4Loop3.java and used as input for Loop2. The following commands are used to execute the program.

*$ javac PharmGKBDrugs4Loop3.java*

*$ java PharmGKBDrugs4Loop3 PharmGKB\_drugs\_loop1\_loop2 OUTPUT\_FILE\_pharmgkbSynonyms OUTPUT\_FILE\_PharmGKBDrugsForLoop3*

The drug concepts that are not concepts or synonyms of UMLS Metathesaurus and DrugBank are filtered from PharmGKB. For each concept, the program collects the synonyms from PharmGKB. DrugsDictionaryCompilerLoop3.java is the related Java program file and can be executed with the following command:

*$ javac DrugsDictionaryCompilerLoop3.java*

*$ java DrugsDictionaryCompilerLoop3 OUTPUT\_FILE\_PharmGKBDrugsForLoop3 OUTPUT\_FILE\_drugsDictionaryLoop3*

Merging output from loop 1, 2 and 3

The following Linux command is used to merge the output from loop 1, 2 and 3.

*$ cat OUTPUT\_FILE\_drugsDictionaryLoop1 OUTPUT\_FILE\_drugsDictionaryLoop2 OUTPUT\_FILE\_drugsDictionaryLoop3 > ChemicalsAndDrugsDictionary*

Lexicon for MedTagger

DrugsDictionaryGenerator.java processes the merged file from loop 1, 2 and 3 to obtain the lexicon in a format that is compatible for use in MedTagger. The following commands are used for executing the program:

*$ javac DrugsDictionaryGenerator.java*

*$ java DrugsDictionaryGenerator ChemicalsAndDrugsDictionary OUTPUT\_FILE\_chemicals\_or\_drugs\_lexicon*

Removal of abbreviations from DrugBank and PharmGKB

Abbreviations across drug entities are noticed to be same in certain drugs. This may lead to false mapping of chemical/drug names. Therefore, we removed the abbreviations <=4 characters from UMLS Metathesaurus (included in UMLS Metathesaurus processing). Similarly, removal of abbreviations from DrugBank and PharmGKB is mandatory. AbbreviationsFromDrugBankAndPharmGKB.java retrieves abbreviations from DrugBank and PharmGKB.

*$ javac AbbreviationsFromDrugBankAndPharmGKB.java*

*$ java AbbreviationsFromDrugBankAndPharmGKB OUTPUT\_FILE\_chemicals\_or\_drugs\_lexicon OUTPUT\_FILE\_abbreviatonsList\_DrugBank\_and\_PharmGKB*

DrugBankAndPharmGKBAbbreviationRemover.java is developed to remove abbreviations <=4 characters from DrugBank and PharmGKB.

*$ javac DrugBankAndPharmGKBAbbreviationRemover.java*

*$ java DrugBankAndPharmGKBAbbreviationRemover OUTPUT\_FILE\_abbreviatonsList\_DrugBank\_and\_PharmGKB OUTPUT\_FILE\_chemicals\_or\_drugs\_lexicon ChemicalsAndDrugsDictionary\_v2.0*

Removal of common English terms as drugs names from PharmGKB

We observed that common English terms do appear as drug names / synonyms in UMLS Metathesaurus. Such names might overlap with the English terms in an input sentences and lead to noise. Therefore, these were removed. We observed the same problem with a few drug names in PharmGKB and removed such terms (given below).

* Cardiovascular System
* Central Nervous System
* Nervous System
* Respiratory System
* Tests for renal function
* Unclassified
* Urine test
* Various
* Washing agents etc.
* Prohibit
* Photodynamic therapy
* Radiotherapy
* Vegetable oil

The updated version of chemicals and drugs dictionary is **ChemicalsAndDrugDictionary\_V3.0\_MedTaggerFormat**