cTAKES 3.2 Dictionaries and Models

Obtaining Prebuilt Dictionaries

The dictionaries and models used during annotation indeed are the cornerstone of quality for your results. The install instructions show you how to get the separately-downloadable ctakes-resources archive (which is not itself released by the Apache Software Foundation) that you need to run most of cTAKES. Those resources include:

- An **RxNorm**_index database (a Lucene index): Contains drug names from RxNorm.
- The **OrangeBook**: If you are not using the drug NER pipeline, the Orange Book is used to filter out what it found in RxNorm so that only things in both RxNorm and Orange Book are annotated. If you use Drug NER, Orange Book filtering is bypassed.
- **UMLS database** (using two hsqldb tables): Contains terms for anatomical sites, procedures, signs/symptoms, and disorders/diseases from SNOMED-CT, NCI Thesaurus, MeSH, and ICD-9 (umls_ms_2011ab) which have been tokenized by cTAKES.
    - 2015 versions
        - SNOMED and RxNorm
        - SNOMED, RxNorm, ICD9, ICD10
- The full **LVG**: From the lexical tools provided by the NLM for word normalization. Used to match similar words, for example the plural and singular forms of a word.

Building Your Own Dictionaries

The UMLS dictionaries within the ctakes-resources archive might not match your underlying data completely. You might require other local terms, etc. To create customized dictionaries for RxNorm, SNOMED-CT, or other vocabularies that are available through the UMLS, you may use one of the dictionary tools currently in development (dictionary-gui and dictionarytool), that can be found in the cTAKES sandbox.

Obtaining Models

As of Apache cTAKES 3.1, the models needed to run cTAKES are included with the convenience binaries.

Building your own Models

You may not need to use any models other than those provided with Apache cTAKES, however they have been trained on a specific set of text (a corpus) which might not match the characteristics of your text. If you want to build or train your own models, please read the cTAKES 3.1 Component Use Guide, particularly:

- Training a sentence detector model
- Training a Part of Speech (POS) tagger model: Building a model - Obtaining training data
- Training a chunker model: Building a model - Prepare GENIA training data
- Training a dependency parser: Training a model - Training data or Training a model in Eclipse