Dataset

In this competition, a specific corpus will be used consisting of a collection of biomedical texts annotated with drug-drug interactions (DDI). This corpus was collected from the database DrugBank.

For each drug, DrugBank contains more than 100 data fields such as drug synonyms, brand names, chemical formula and structure, drug categories, ATC and AHFS codes (codes of standard drug families), mechanism of action, indication, dosage forms, toxicity, among other ones. In addition, DrugBank contains a link to a document describing DDIs in unstructured text for each drug. We have used these documents field as a source of unstructured textual information on DDIs.

Once the recopilation of data from Drugbank is done, the documents were processed by the MetaMap tool to syntactically and semantically analyze the documents in the corpus. MMTx performs sentence splitting, tokenization, POS-tagging, shallow syntactic parsing, and linking of phrases with UMLS concepts. In particular, MMTx allows us to recognize and annotate a variety of biomedical entities occurring in texts according to the UMLS semantic types.

An experienced pharmacist reviewed the semantic annotation provided by MMTx and recommended the inclusion of the following UMLS semantic types as possible types of interacting drugs.

- Clinical Drug (clnd)
- Pharmacological Substance (phsu)
- Antibiotic (antb)
- Biologically Active Substance (bacs)
- Chemical Viewed Structurally (chvs)
- Amino Acid, Peptide, or Protein (aapp)

The main value of the DrugDDI corpus comes from its annotation since all the documents have been marked-up with drug-drug interactions

by a researcher with pharmaceutical background and a pharmacist.

Since most of the existing approaches for relation extraction usually assume that the argument entities of the relation occur in the same sentence, we only considered the interactions between drugs within the same sentence.

Although there may be relations between drugs in different sentences, they have not been annotated in the DrugDDI corpus.

We provide our corpus in two different formats: (1) a format based on the information provided by MMTx and (2) the unified XML format for Protein-Protein Interaction Extraction proposed in [1]. Hence, participants should choose between these two formats depending on their preferences, since some systems may have no use for MMTx information.

One of the problems with annotated datasets is that they can come in different formats, thus, we have also generated our corpus using the unified XML format for PPIE.

Drug names were automatically annotated by using the MMTx tool. Unfortunately, this tool makes some mistakes.

As source of unstructured textual information on drugs and their interactions, we have used the DrugBank database. This database offers a complete collection of DDIs, which was compiled from several resources, checked by an accredited pharmacist and entered manually. This collection consists of 714 food-interactions and 13,242 drug-drug interactions. The interactions are contained in the structured information fields "Food Interactions" and "Drug Interactions".

In a previous version of DrugBank, additional information could be found in the field "Interactions". This field contained a link to a document describing DDIs in unstructured text. This document did not only contain a detailed description on the interactions contained in the above structured fields, but also offers information on other interactions which have not collected in them. We used the "Interactions" field as a source of unstructured textual information on DDIs. Unfortunately, this field is not available for the current version of DrugBank.

The average number of drugs per document is 25.88, and the average number of drugs per sentence is 2.63. The average number of interactions per document is 5.52 and per sentence 0.56.

The corpus was split into in order to build the training and testing datasets. The training dataset necessary for the task will be available in April 1st and contains a total of 2812 sentences that contain two or more drugs, although only 1530 contain at least one interaction.

A total of 2402 drug-drug interactions and 11260 drugs have been identified in the training dataset. The test dataset necessary for the evaluation part of the task will be available in May 30th.

[1] S. Pyysalo, A. Airola, J. Heimonen, J. Bjorne, F. Ginter, and T. Salakoski. Comparative analysis of protein-protein interaction corpora. BMC Bioinformatics