Task 9.1: Recognition and Classification of drug names.

This task concerns the named entity extraction of mentions of pharmacological substances in text. This named entity task is a crucial first step for information extraction of drug-drug interactions. There are four general types of entities:

1. Drug: any chemical agent used in the treatment, cure, prevention or diagnosis of diseases which has been approved for human use. This type only represents generic drugs. "A generic drug is a pharmaceutical product, usually intended to be interchangeable with an innovator product, that is manufactured without a license from the innovator company and marketed after the expiry date of the patent or other exclusive rights."[1]
2. Brand: any drug that was first developed by a pharmaceutical company.
3. Group: any term in text designating a chemical or pharmacologic relationship among a group of drugs
4. No-Human: any chemical agent that affects living organisms. It's an active substance but it has not been approved to be used in humans with a medical purpose

Other types of entities (e.g. cells, food, etc) are out of our scope. **Please follow our guidelines.**

The following figure illustrates the format of the training data set.

```
−<document id="DrugDDI.d505" origId="Abarelix">
 −<sentence id="DrugDDI.d505.s0" origId="s0" text="No formal drug/drug interaction studies with Plenaxis were performed.">
    <entity id="DrugDDI.d505.s0.e0" origId="" charOffset="45-52" type="brand" text="Plenaxis"/>
  </sentence>
 −<sentence id="DrugDDI.d505.s1" origId="s1" text="Cytochrome P-450 is not known to be involved in the metabolism of Plenaxis.">
    <entity id="DrugDDI.d505.s1.e0" origId="" charOffset="66-73" type="brand" text="Plenaxis"/>
  </sentence>
 −<sentence id="DrugDDI.d505.s2" origId="s2" text="Plenaxis is highly bound to plasma proteins (96 to 99%).">
    <entity id="DrugDDI.d505.s2.e0" origId="" charOffset="0-7" type="brand" text="Plenaxis"/>
  </sentence>
 −<sentence id="DrugDDI.d505.s3" origId="s3" text="Laboratory Tests Response to Plenaxis should be monitored by measuring
  serum total testosterone concentrations just prior to administration on Day 29 and every 8 weeks thereafter.">
    <entity id="DrugDDI.d505.s3.e0" origId="" charOffset="29-36" type="brand" text="Plenaxis"/>
    <entity id="DrugDDI.d505.s3.e1" origId="s3.p30" charOffset="83-94" type="drug" text="testosterone"/>
  </sentence>
 −<sentence id="DrugDDI.d505.s4" origId="s4" text="Serum transaminase levels should be obtained before starting treatment with
  Plenaxis and periodically during treatment.">
    <entity id="DrugDDI.d505.s4.e0" origId="" charOffset="76-83" type="brand" text="Plenaxis"/>
  </sentence>
  <sentence id="DrugDDI.d505.s5" origId="s5" text="Periodic measurement of serum PSA levels may also be considered."/>
</document>
```

Figure 1 Format of the training dataset for the Drug Name Recognition task.

Output Format

For evaluation, a held-out part of the same corpus, consisting of 52 documents from DrugBank and 57 MedLine abstracts, will be provided with the gold annotation hidden. The goal for

[1] http://www.who.int/trade/glossary/story034/en/index.html

participating systems is to recreate the gold annotation. Participant systems will be required to return the mention, the type and the start and end indices corresponding to all the pharmacological substances in a given sentence. Each participant system must output an ASCII list of reported drug mentions, one per line, and formatted as:

IdSentence|startOffset1-endOffset1; startOffset2-endOffset2;…|mention|type

where

- **IdSentence** is from the sentence of the mention, e.g, DrugDDI.d505.s3
- **startOffset1-endOffset1**: startOffset1is the position of the first character of the mention in the sentence (0 is the first character of the sentence) while endOffset1 is the position of the last character of the mention in the sentence. For example, for the entity DrugDDI.d505.s3.e0, its startOffset is 29 and its endOffset is 36.
- When the mention is a discontinuous name, this may contain the start and end positions of all parts of the mention separated by semicolon (see Figure 1.2, entity s5.e4)
- **mention** corresponds to the text of the pharmacological substance in the sentence.
- **type** corresponds to the type of the pharmacological substance (drug, brand, group, drug-n). Participants may return null for this attribute.

Multiple mentions from the same entity should appear on separate lines. A sentence is not required to have any mentions.

```
−<sentence id="DrugDDI.d42.s5" origId="s5" text="If a patient requires TIKOSYN and anti-ulcer therapy, it
 is suggested that omeprazole, ranitidine, or antacids (aluminum and magnesium hydroxides) be used as
 alternatives to cimetidine, as these agents have no effect on the pharmacokinetic profile of TIKOSYN.">
   <entity id="DrugDDI.d42.s5.e0" origId="s5.p47" charOffset="22-28" type="brand" text="TIKOSYN"/>
   <entity id="DrugDDI.d42.s5.e1" origId="s5.p53" charOffset="75-84" type="drug" text="omeprazole"/>
   <entity id="DrugDDI.d42.s5.e2" origId="s5.p54" charOffset="87-96" type="drug" text="ranitidine"/>
   <entity id="DrugDDI.d42.s5.e3" origId="s5.p56" charOffset="102-109" type="group" text="antacids"/>
   <entity id="DrugDDI.d42.s5.e4" origId="" charOffset="112-119;135-143" type="drug" text="aluminum
   hydroxide"/>
   <entity id="DrugDDI.d42.s5.e5" origId="s5.p60" charOffset="125-143" type="drug" text="magnesium
   hydroxide"/>
   <entity id="DrugDDI.d42.s5.e6" origId="s5.p65" charOffset="174-183" type="drug"
   text="cimetidine"/>
   <entity id="DrugDDI.d42.s5.e7" origId="s5.p71" charOffset="251-257" type="brand"
   text="TIKOSYN"/>
   <ddi id="DrugDDI.d42.s5.d0" e1="DrugDDI.d42.s5.e1" e2="DrugDDI.d42.s5.e6" type="advise"/>
 </sentence>
```

Figure 2 Example of discontinuous name

The bellow example contains the list of drug mentions for the Figure 1.2:

DrugDDI.d42.s5|22-28|TIKOSYN|brand
DrugDDI.d42.s5|75-84|omeprazole|drug
DrugDDI.d42.s5|87-96|ranitidine|drug
DrugDDI.d42.s5|102-109|antacids|group
DrugDDI.d42.s5|112-119;135-143|aluminum hydroxide|drug
DrugDDI.d42.s5|125-143|magnesium hydroxide|drug
DrugDDI.d42.s5|174-183|cimetidine|drug

Up to three runs may be submitted by each team. Details about submission procedures will be communicated through our web site. A script will be made available to ensure that submission files comply with the prescribed format.

Evaluation

System performance will be scored automatically by how well the generated pharmacological substance list corresponds to one generated by human annotators. A named entity is correct only if it is an exact match of the corresponding gold standard entity in the data, that is, their type, start and end positions are the same.

Evaluation results are reported using the standard precision/recall/f-score metrics. Precision is the percentage of named entities found by the learning system that are correct. Recall is the percentage of named entities present in the corpus that are found by the system. F-score is the harmonic mean of precision and recall:

$$F1 = \frac{2 * precision * recall}{(precision + recall)}$$

These metrics will be computed for each type of entity (drug, brand, group, no-human) and overall.

File naming conventions

All files in the task follow the same naming convention:

task9.1_GROUP_RUN.txt

where:

- GROUP is necessary to identify which group made the submission.
- RUN is a integer value between 1 and 3, to distinguish multiple submissions.

For example, the UC3M team may submit the following files for the NER subtask: task9.1_UC3M _1.txt, task9.1_UC3M _2.txt, task9.1_UC3M _3.txt.