Evaluation of the SemEval-2013 Task 9.2: Extraction of Drug Drug Interactions

Contenido

1. Introduction

The goal of this subtask is the extraction of drug-drug interactions from biomedical texts. However, while the previous DDIExtraction 2011 task focused on the identification of all possible pair of interacting drugs, DDIExtraction 2013 also pursues the classification of each drug-drug interaction according to one of the four following types: advise, effect, mechanism, int. A detailed description of these types can find in our annotation guidelines.

Please, note the Named entity recognition is not necessary to address the task because gold standard annotations (correct, human-created annotations) of pharmacological substances are provided to participants both for training and test data.

Participant systems will be required to return a list that includes all pairs of drugs in each sentence and its prediction. Each participant system must output an ASCII list including all pairs of drugs in each sentence, one per line (multiple ddis from the same sentence should appear on separate lines), its prediction (1 if the pair is a DDI and 0 e.o.c) and its type (label **null** when the prediction value is 0), and formatted as:

IdSentence|IdDrug1|IdDrug2|prediction|type.

2. Evaluation

Evaluation is relation-oriented and based on the standard precision, recall and F-score metrics. Note that only relations are evaluated since entities will be included in the test dataset.

In our task, we will evaluate the results of the participating systems according to several evaluation criteria:

1) Strict evaluation: a DDI is correctly detected only if the system is able to assign the correct prediction label and the correct type to it. In other words, a pair is correct only if both prediction and type are correct. When prediction is 0, type may be empty or null.
2) Partial evaluation: a pair is correct when if its prediction label matches in the gold annotation

Evaluation results will be reported using the standard precision/recall/f-score metrics:

- Precision is the percentage of DDIs found by the learning system that are correct. That is, precision is the ratio between the number of DDIs correctly detected (true positives) and the total number of DDIs that were found by the system (true positives + false positives).

$$P = \frac{TP}{TP + FP}$$

- Recall is the percentage of DDIs present in the corpus that are found by the system. In other words, recall is the ratio between the number of DDIs correctly detected (true positives) and the total number of drug entities in the gold standard (true positives + false negatives).

$$R = \frac{TP}{TP + FN}$$

- F-score is the harmonic mean of precision and recall:

$$F1 = \frac{2 * P * R}{P + R}$$

The classification of DDIs will also be evaluated to assess which drug interactions are most difficult to detect. We will calculate precision, recall and f-measure for each DDI type and then their macro-average measures will be provided. While micro-averaged F-scoreis calculated by constructing a global contingency table and then calculating precision and recall, macro-averaged F-score is calculated by first calculating precision and recall for each type and then taking the average of these. Thus, for example, the precision for *mechanism* relationshipscan be defined as the ratio between the number of DDIs correctly classified as *mechanism* and the total number of DDIs that were classified as *mechanism* (including the ones wrongly assignedto this type). Similarly the recall for *mechanism* relationships is defined as the ratio between the number of DDIs correctly classified as *mechanism* and the total number of DDIs with *mechanism* type in the gold standard. The precision and recall for the rest ofDDIs types is defined in a similar manner.

Evaluating each DDI type separately will allow us to assess the level of difficulty of detecting each type of interaction. Additionally, it is important to note that the scores achieved on the most frequent DDI type have a much greater impact on overall performance than those achieved on the DDI types with few instances. Therefore, by calculating scores for each type of DDI, we can better assess the performance of the algorithms proposed by the participating systems.