Evaluation of the SemEval-2013 Task 9.1:

Recognition and Classification of pharmacologicalsubstances

Contenido

1. Introduction

The task of recognition and classification of drug names 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. Theannotationschemacontainsfourtypes 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.
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. drug_n: 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.

This document describes the methodology that will be used to evaluate the performance of the participating systems in task 9.1. Recognition and classification of pharmacological substances is a named entity recognition and classification (NERC) task. System performance should be scored automatically by how well the generated pharmacological substance list corresponds to one generated by human annotators.

2. Evaluation Metrics

The evaluation of a system requires comparing its outputs with the gold-standard annotations. The main types of errors of a NERC system are described in the example bellow.

Figure 1 shows the gold-standard annotation of a sentence. Let us hypothesize that Table 1 shows the answer provided by a participating system (Figure 2 shows this ouput in XML format).

```
- <sentence id="DDI-DrugBank.d558.s33" text="In addition, studies in healthy volunteers have shown that TIKOSYN does not
  affect the pharmacokinetics or pharmacodynamics of warfarin, or the pharmacokinetics of propranolol (40 mg twice daily),
  phenytoin, theophylline, or oral contraceptives.">
    <entity id="DDI-DrugBank.d558.s33.e0" charOffset="59-65" type="brand" text="TIKOSYN"/>
    <entity id="DDI-DrugBank.d558.s33.e1" charOffset="127-134" type="drug" text="warfarin"/>
    <entity id="DDI-DrugBank.d558.s33.e2" charOffset="164-174" type="drug" text="propranolol"/>
    <entity id="DDI-DrugBank.d558.s33.e3" charOffset="197-205" type="drug" text="phenytoin"/>
    <entity id="DDI-DrugBank.d558.s33.e4" charOffset="208-219" type="drug" text="theophylline"/>
    <entity id="DDI-DrugBank.d558.s33.e5" charOffset="230-243" type="group" text="contraceptives"/>
  </sentence>
```

Figure 1 Example of a gold-standard sentence

ID sentence|startOffset-endOffset|span|type
DDI-DrugBank.d558.s33|24-30[1]|healthy|brand
DDI-DrugBank.d558.s33|124-134|of warfarin|drug
DDI-DrugBank.d558.s33|164-174|propranolol|brand
DDI-DrugBank.d558.s33|197-205|phenytoin|drug
DDI-DrugBank.d558.s33|208-219|theophylline|drug
DDI-DrugBank.d558.s33|225-243|oral contraceptives|drug

Table 1 Output format proposed by the task 9.1

```
- <sentence id="DDI-DrugBank.d558.s33" text="In addition, studies in healthy volunteers have shown that TIKOSYN does not
  affect the pharmacokinetics or pharmacodynamics of warfarin, or the pharmacokinetics of propranolol (40 mg twice daily),
  phenytoin, theophylline, or oral contraceptives.">
    <entity id="DDI-DrugBank.d558.s33.e0" charOffset="24-30" type="brand" text="healthy"/>
    <entity id="DDI-DrugBank.d558.s33.e1" charOffset="124-134" type="drug" text="of warfarin"/>
    <entity id="DDI-DrugBank.d558.s33.e2" charOffset="164-174" type="brand" text="propranolol"/>
    <entity id="DDI-DrugBank.d558.s33.e3" charOffset="197-205" type="drug" text="phenytoin"/>
    <entity id="DDI-DrugBank.d558.s33.e4" charOffset="208-219" type="drug" text="theophylline"/>
    <entity id="DDI-DrugBank.d558.s33.e5" charOffset="225-243" type="drug" text="oral contraceptives"/>
  </sentence>
```

Figure 2 Example of the system's output in XML format

Note that only two entities were correctly labeled by the system: phenytoin (e3) and theophylline (e4). Therefore, the system produces five errors that are explained bellow:

i. The system labels an entity that does not exist: healthy (e0).
ii. The system is not able to recognize an entity found in the gold-standard annotation: TIKOSYN (e0 in the gold-standard).
iii. The system labels an entity with wrong boundaries: "of warfarin" (e1)
iv. An entity is correctly recognized but the system labels it a wrong type: propranolol (e2).
v. The system fails to establish both type and boundaries: "oral contraceptives" (e5). Although "oral contraceptives" is a well-known group, the annotation guidelines[2] stated that the routes of administrationshouldnot be included in the annotation.

The major forums of the NERC research community (such as MUC-7[3], CoNLL 2003[4] or ACE 2007[5]) have proposed several techniques to assess the performance of NERC systems. However, these techniques can vary markedly in terms of their scores [1].

ACE defined a set of metrics which allow assigning a parameterized weight for each entity type, customizing the cost of error (COST) and dealing with various evaluation issues such as partial match and wrong type. However, ACE evaluation is very complex because the final scores are only comparable when weights are fixed. Moreover, the error analysis is very hard to understand because the scores are not intuitive.

[1] If the mention has a discontinuous name, this ouput should contain the start and end positions of all parts of the mention separated by semicolon (eg, 12-15;21-25)

[2] http://www.cs.york.ac.uk/semeval-2013/task9/data/uploads/annotation_guidelines_ddi_corpus.pdf
[3] http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_proceedings/overview.html
[4] http://www.cnts.ua.ac.be/conll2003/ner/
[5] http://www.itl.nist.gov/iad/mig/tests/ace/ace07/

MUC and CoNLL 2003 used the standard precision/recall/f-score metrics to compare their partipating systems. Also, the main shared tasks (BioNLP 2004[6], BioCreative I, II and IV[7]) in the biomedical domain have continued using these metrics to evaluate the outputs of their participating teams.

CoNLL 2003 proposes an exact-match evaluation, in which an entity is correct only if it is an exact match of the corresponding gold standard entity in the data. However, requiring exact matches may be too restrictive and lead to a loss of relevant information for many applications, for example, in the biomedical domain. Other tasks have followed more relaxed matching criteria such as right/left boundary matching (if a boundary matches -either right or left-exactly the entity is scored as correct), approximate matching (the tagged entity is a substring of the gold entity or vice versa) or partial matching (tagged entity and gold entity have some overlapping text) [2].

In our task, we will evaluate the results of the participating systems according to several evaluation criteria. Firstly, we propose a strict evaluation, which does not only demand exact boundary match, but also requires that both mentions have the same entity type.We are aware that this strict criterion might be too restrictive for our overall goal (extraction of drug interactions) because misses partial matches, which can provide useful information for a DDI extraction system. Our evaluation metrics should score if a system is able to identify the exact span of an entity (regardless of the type) andif it is able to assign the correct entity type (regardless of the boundaries). Thus, our evaluation script will output four sets of scores according to:

1) Strict evaluation,
2) Exact boundary matching (regardless to the type),
3) Partial boundary matching (regardless to the type).
4) Type matching (some overlap between the tagged entity and the gold entitity is required)

In order to calculate precision and recall, we will use the the scoring categories[8] proposed by MUC:

- COR: the system's output and the gold-standard annotation agree.
- INC: the system's output and the gold-standard annotation disagree.
- PAR: the system's output and the gold-standard annotation are not identical but have some overlapping text. This category makes sense only if the partial match is allowed.
- MIS: there is a gold standard entity that is not identified by the system
- SPU: the system labels an entity that does not exist in the gold-standard.

For both the boundaries and the type, the following measures are calculated:

- The number of correct answers: *COR*

[6] http://www.nactem.ac.uk/tsujii/GENIA/ERtask/report.html
[7] http://www.biocreative.org/tasks/
[8] http://www.itl.nist.gov/iaui/894.02/related_projects/muc/muc_sw/muc_sw_manual.html

- The number of annotations in the gold-standard which contribute to the final score:

$$POSSIBLE\ (POS) = COR + INC + PAR + MIS = TP + FN$$

- The total number of annotations produced by the system:

$$ACTUAL\ (ACT) = COR + INC + PAR + SPU = TP + FP$$

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

- Precision is the percentage of named entities found by the learning system that are correct. If exact match is required then the formula to calculate precision is:

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

  The formula to rating the partially correct answers is:

$$P = \frac{COR + 0.5 * PAR}{ACT}$$

- Recall is the percentage of named entities present in the corpus that are found by the system. If exact match is required then the formula to calculate recall is:

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

  The formula to rating the partially correct answers is:

$$R = \frac{COR + 0.5 * PAR}{POS}$$

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

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

These metrics are calculated over all entities and on both axes (type and text) in order to evaluate the performance of each axe separately. The final score is the micro-averaged F-measure, which is calculated over all entity types without distinction. The main advantage of the micro-average F1 is that takes into account all possible type of errors made by a NERC system. The following table showsthe scoresfor the previous example shown in see Table 1:

Measure	Strict	Exact Matching	Partial Maching	Type
COR	2	3	3	3
INC	3	2	0	2
PAR	0	0	2	0
MIS	1	1	1	1
SPU	1	1	1	1
Precision	0.33	0.5	0.5	0.5
Recall	0.33	0.5	0.67	0.5
F1	0.33	0.5	0.57	0.5

Table 2 Scores for the example shown in Table 1 (see also Figure 2)

type (gold standard)	span (gold-standard)	type (answer)	span (answer)	Type Matching	Partial Matching	Exact Matching	Strict
brand	TIKOSYN			MIS	MIS	MIS	MIS
		brand	healthy	SPU	SPU	SPU	SPU
drug	warfarin	drug	of warfarin	COR	PAR	INC	INC
drug	propranolol	brand	propranolol	INC	COR	COR	INC
drug	phenytoin	drug	phenytoin	COR	COR	COR	COR
Drug	theophylline	drug	theophylline	COR	COR	COR	COR
group	contraceptives	drug	oral contraceptives	INC	PAR	INC	INC

Table 3 Alignment between sentences shown in Figure 1 and Figure 2

Additionally, we will calculate precision, recall and f-measure for each type of entity and then their macro-average measures will be provided.Calculating these metrics for each entity type will allow us to evaluate the level of difficulty of recognizing each entity type. In addition to this, since not all entity types have the same frecuency, we can better assess the performance of the algorithms proposed by the participating systems. This is mainly because the results achieved on the most frequent entity type have a much greater impact on overall performance than those obtaiened on the entity types with few instances.

Thus, for example, the precision for drug entities can be defined as the ratio between the number of entities correctly classified as drug and the total number of entities that were classified as drug (including the ones wrongly assignedto this type). Similarly the recall for drug entities is defined as the ratio between the number of entities correctly classified as drug and the total number of drug entities in the gold standard. The precision and recall for the rest of entity types is defined in a similar manner.

References

[1]. Nadeau, David and Sekine, Satoshi. **A survey of named entity recognition and classification.** Lingvisticae Investigationes 2007;30(1):3-26.

[2]. Tsai, Richard Tzong-Han and Wu, Shih-Hung and Chou, Wen-Chi and Lin, Yu-Chun and He, Ding and Hsiang, Jieh and Sung, Ting-Yi and Hsu, Wen-Lian. **Various criteria in the evaluation of biomedical named entity recognition**. BMC bioinformatics 2006;7(1):92.