SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013)

Isabel Segura-Bedmar, Paloma Martínez, María Herrero-Zazo
Universidad Carlos III de Madrid
Av. Universidad, 30, Leganés 28911, Spain
{isegura,pmf}@inf.uc3m.es, mhzazo@pa.uc3m.es

Abstract

The DDIExtraction 2013 task concerns the recognition of drugs and extraction of drug-drug interactions that appear in biomedical literature. We propose two subtasks for the DDIExtraction 2013 Shared Task challenge: 1) the recognition and classification of drug names and 2) the extraction and classification of their interactions. Both subtasks have been very successful in participation and results. There were 14 teams who submitted a total of 38 runs. The best result reported for the first subtask was F1 of 71.5% and 65.1% for the second one.

1 Introduction

The definition of drug-drug interaction (DDI) is broadly described as a change in the effects of one drug by the presence of another drug (Baxter and Stockely, 2010). The detection of DDIs is an important research area in patient safety since these interactions can become very dangerous and increase health care costs. Drug interactions are frequently reported in journals, making medical literature the most effective source for their detection (Aronson, 2007). Therefore, Information Extraction (IE) can be of great benefit in the pharmaceutical industry allowing identification and extraction of relevant information on DDIs and providing an interesting way of reducing the time spent by health care professionals on reviewing the literature.

The DDIExtraction 2013 follows up on a first event organized in 2011, DDIExtraction 2011 (Segura-Bedmar et al., 2011b) whose main goal was the detection of drug-drug interactions from biomedical texts. The new edition includes in addition to DDI extraction also a supporting task, the recognition and classification of pharmacological substances. DDIExtraction 2013 is designed to address the extraction of DDIs as a whole, but divided into two subtasks to allow separate evaluation of the performance for different aspects of the problem. The shared task includes two challenges:

- Task 9.1: Recognition and classification of pharmacological substances.
- Task 9.2: Extraction of drug-drug interactions.

Additionally, while the datasets used for the DDIExtraction 2011 task were composed by texts describing DDIs from the DrugBank database(Wishart et al., 2006), the new datasets for DDIExtraction 2013 also include MedLine abstracts in order to deal with different types of texts and language styles.

This shared task has been conceived with a dual objective: advancing the state-of-the-art of text-mining techniques applied to the pharmacological domain, and providing a common framework for evaluation of the participating systems and other researchers interested in the task.

In the next section we describe the DDI corpus used in this task. Sections 3 and 4 focus on the description of the task 9.1 and 9.2 respectively. Finally, Section 5 draws the conclusions and future work.

2 The DDI Corpus

The DDIExtraction 2013 task relies on the DDI corpus, which is a semantically annotated corpus of

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, June 14-15, 2013.

documents describing drug-drug interactions from the DrugBank database and MedLine abstracts on the subject of drug-drug interactions.

The DDI corpus consists of 1,017 texts (784 DrugBank texts and 233 MedLine abstracts) and was manually annotated with a total of 18,491 pharmacological substances and 5,021 drug-drug interactions (see Table 1). A detailed description of the method used to collect and process documents can be found in (Segura-Bedmar et al., 2011a). The corpus is distributed in XML documents following the unified format for PPI corpora proposed by Pyysalo et al., (2008) (see Figure 1). A detailed description and analysis of the DDI corpus and its methodology are included in an article currently under review by BioInformatics journal.[1]

The corpus was split in order to build the datasets for the training and evaluation of the different participating systems. Approximately 77% of the DDI corpus documents were randomly selected for the training dataset and the remaining (142 DrugBank texts and 91 MedLine abstracts) was used for the test dataset. The training dataset is the same for both subtasks since it contains entity and DDI annotations. The test dataset for the task 9.1 was formed by discarding documents which contained DDI annotations. Entity annotations were removed from this dataset to be used by participants. The remaining documents (that is, those containing some interaction) were used to create the test dataset for task 9.2. Since entity annotations are not removed from these documents, the test dataset for the task 9.2 can also be used as additional training data for the task 9.1.

3 Task 9.1: Recognition and classification of pharmacological substances.

This task concerns the named entity extraction of pharmacological substances in text. This named entity task is a crucial first step for information extraction of drug-drug interactions. In this task, four types of pharmacological substances are defined: *drug* (generic drug names), *brand* (branded drug names), *group* (drug group names) and *drug-n* (active substances not approved for human use). For a more detailed description, the reader is directed to our annotation guidelines.[2]

		Training	Test for task 9.1	Test for task 9.2
DDI-DrugBank	documents	572	54	158
	sentences	5675	145	973
	drug	8197	180	1518
	group	3206	65	626
	brand	1423	53	347
	drug_n	103	5	21
	mechanism	1260	0	279
	effect	1548	0	301
	advice	819	0	215
	int	178	0	94
DDI-MedLine	documents	142	58	33
	sentences	1301	520	326
	drug	1228	171	346
	group	193	90	41
	brand	14	6	22
	drug_n	401	115	119
	mechanism	62	0	24
	effect	152	0	62
	advice	8	0	7
	int	10	0	2

Table 1: Basic statistics on the DDI corpus.

For evaluation, a part of the DDI corpus consisting of 52 documents from DrugBank and 58 MedLine abstracts, is provided with the gold annotation hidden. The goal for participating systems is to recreate the gold annotation. Each participant system must output an ASCII list of reported entities, one per line, and formatted as:

IdSentence|startOffset-endOffset|text|type

Thus, for each recognized entity, each line must contain the id of the sentence where this entity appears, the position of the first character and the one of the last character of the entity in the sentence, the text of the entity, and its type. When the entity is a discontinuous name (eg. *aluminum* and magnesium *hydroxide*), this second field must contain the start and end positions of all parts of the entity separated by semicolon. Multiple mentions from the same sentence should appear on separate lines.

3.1 Evaluation Metrics

This section describes the methodology that is used to evaluate the performance of the participating systems in task 9.1.

The major forums of the Named Entity Recognition and Classification (NERC) research community (such as MUC-7 (Chinchor and Robinson, 1997), CoNLL 2003 (Tjong Kim Sang and De Meulder, 2003) or ACE07 have proposed several techniques to assess the performance of NERC systems. While

[1] M. Herrero-Zazo, I. Segura-Bedmar, P. Martínez. 2013. The DDI Corpus: an annotated corpus with pharmacological substances and drug-drug interactions, submitted to BioInformatics

[2] http://www.cs.york.ac.uk/semeval-2013/task9/

```
−<document id="DDI-DrugBank.d372">
 −<sentence id="DDI-DrugBank.d372.s0" text="Cytadren accelerates the metabolism of dexamethasone;">
    <entity id="DDI-DrugBank.d372.s0.e0" charOffset="0-7" type="brand" text="Cytadren"/>
    <entity id="DDI-DrugBank.d372.s0.e1" charOffset="39-51" type="drug" text="dexamethasone"/>
    <pair id="DDI-DrugBank.d372.s0.p0" e1="DDI-DrugBank.d372.s0.e0" e2="DDI-DrugBank.d372.s0.e1" ddi="true" type="mechanism"/>
  </sentence>
 −<sentence id="DDI-DrugBank.d372.s1" text="therefore, if glucocorticoid replacement is needed, hydrocortisone should be prescribed.">
    <entity id="DDI-DrugBank.d372.s1.e0" charOffset="14-27" type="group" text="glucocorticoid"/>
    <entity id="DDI-DrugBank.d372.s1.e1" charOffset="52-65" type="drug" text="hydrocortisone"/>
    <pair id="DDI-DrugBank.d372.s1.p0" e1="DDI-DrugBank.d372.s1.e0" e2="DDI-DrugBank.d372.s1.e1" ddi="false"/>
  </sentence>
 −<sentence id="DDI-DrugBank.d372.s2" text="Aminoglutethimide diminishes the effect of coumarin and warfarin.">
    <entity id="DDI-DrugBank.d372.s2.e0" charOffset="0-16" type="drug" text="Aminoglutethimide"/>
    <entity id="DDI-DrugBank.d372.s2.e1" charOffset="43-50" type="group" text="coumarin"/>
    <entity id="DDI-DrugBank.d372.s2.e2" charOffset="56-63" type="drug" text="warfarin"/>
    <pair id="DDI-DrugBank.d372.s2.p0" e1="DDI-DrugBank.d372.s2.e0" e2="DDI-DrugBank.d372.s2.e1" ddi="true" type="effect"/>
    <pair id="DDI-DrugBank.d372.s2.p1" e1="DDI-DrugBank.d372.s2.e0" e2="DDI-DrugBank.d372.s2.e2" ddi="true" type="effect"/>
    <pair id="DDI-DrugBank.d372.s2.p2" e1="DDI-DrugBank.d372.s2.e1" e2="DDI-DrugBank.d372.s2.e2" ddi="false"/>
  </sentence>
</document>
```


Figure 1: Example of an annotated document of the DDI corpus.

	Team	Affiliation	Description
Task 9.1	LASIGE(Grego et al., 2013)	University of Lisbon, Portugal	Conditional random fields
	NLM_LHC	National Library of Medicine, USA	Dictionary-based approach
	UEM_UC3M(Sanchez-Cisneros and Aparicio, 2013)	European U. of Madrid, Carlos III University of Madrid, Spain	Ontology-based approach
	UMCC_DLSI(Collazo et al., 2013)	Matanzas University, Cuba	j48 classifier
	UTurku(Björne et al., 2013)	University of Turku, Finland	SVM classifier (TEES system)
	WBI_NER(Rocktäschel et al., 2013)	Humboldt University of Berlin, Germany	Conditional random fields
Task 9.2	FBK-irst (Chowdhury and Lavelli, 2013c)	FBK-irst, Italy	hybrid kernel + scope of negations and semantic roles
	NIL_UCM(Bokharaeian, 2013)	Complutense University of Madrid, Spain	SVM classifier (Weka SMO)
	SCAI(Bobić et al., 2013)	Fraunhofer SCAI, Germany	SVM classifier (LibLINEAR)
	UC3M(Sanchez-Cisneros, 2013)	Carlos III University of Madrid, Spain	Shallow Linguistic Kernel
	UCOLORADO_SOM(Hailu et al., 2013)	University of Colorado School of Medicine, USA	SVM classifier (LIBSVM)
	UTurku(Björne et al., 2013)	University of Turku, Finland	SVM classifier (TEES system)
	UWM-TRIADS(Rastegar-Mojarad et al., 2013)	University of Wisconsin-Milwaukee, USA	Two-stage SVM
	WBI_DDI(Thomas et al., 2013)	Humboldt University of Berlin, Germany	Ensemble of SVMs

Table 2: Short description of the teams.

ACE evaluation is very complex because its scores are not intuitive, MUC and CoNLL 2003 used the standard precision/recall/f-score metrics to compare their participating systems. The main shared tasks in the biomedical domain have continued using these metrics to evaluate the outputs of their participant teams.

System performance should be scored automatically by how well the generated pharmacological substance list corresponds to the gold-standard annotations. In our task, we 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 may be too restrictive for our overall goal (extraction of drug interactions) because it 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) and if 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 (exact-boundary and type matching).
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).

Evaluation results are reported using the standard precision/recall/f-score metrics. We refer the reader to (Chinchor and Sundheim, 1993) for a more detailed description of these metrics.

These metrics are calculated over all entities and on both axes (type and span) 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 it

takes into account all possible types of errors made by a NERC system.

Additionally, we calculate precision, recall and f-measure for each entity type and then their macro-average measures are provided. Calculating these metrics for each entity type allows us to evaluate the level of difficulty of recognizing each entity type. In addition to this, since not all entity types have the same frequency, 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 obtained on the entity types with few instances.

3.2 Results and Discussion

Participants could send a maximum of three system runs. After downloading the test datasets, they had a maximum of two weeks to upload the results. A total of 6 teams participated, submitting 16 system runs. Table 2 lists the teams, their affiliations and a brief description of their approaches. Due to the lack of space we cannot describe them in this paper. Tables 3, 4 and 5 show the F1 scores for each run in alphabetic order. The reader can find the full ranking information on the SemEval-2013 Task 9 website[3].

The best results were achieved by the WBI team with a conditional random field. They employed a domain-independent feature set along with features generated from the output of ChemSpot (Rocktäschel et al., 2012), an existing chemical named entity recognition tool, as well as a collection of domain-specific resources. Its model was trained on the training dataset as well as on entities of the test dataset for task 9.2. The second top best performing team developed a dictionary-based approach combining biomedical resources such as DrugBank, the ATC classification system,[4] or MeSH,[5] among others. Regarding the classification of each entity type, we observed that brand drugs were easier to recognize than the other types. This could be due to the fact that when a drug is marketed by a pharmaceutical company, its brand name is carefully selected to be short, unique and easy to remember (Boring, 1997). On the other hand, substances not approved for human use (*drug-n*) were more difficult, due to the greater variation and complexity in their naming. In fact, the UEM_UC3M team was the only team who obtained an F1 measure greater than 0 on the DDI-DrugBank dataset. Also, this may indicate that this type is less clearly defined than the others in the annotation guidelines. Another possible reason is that the presence of such substances in this dataset is very scarce (less than 1%). It is interesting that almost every participating system was better in detecting and classifying entities of a particular class compared to all other systems. For instance, on the whole dataset the dictionary-based system from NLM_LHC had it strengths at *drug* entities, UEM_UC3M at *drug_N* entities, UTurku at *brand* entities and WBI_NER at *group* entities.

Finally, the results on the DDI-DrugBank dataset are much better than those obtained on the DDI-MedLine dataset. While DDI-DrugBank texts focus on the description of drugs and their interactions, the main topic of DDI-MedLine texts would not necessarily be on DDIs. Coupled with this, it is not always trivial to distinguish between substances that should be classified as pharmacological substances and those who should not. This is due to the ambiguity of some pharmacological terms. For example, *insulin* is a hormone produced by the pancreas, but can also be synthesized in the laboratory and used as drug to treat insulin-dependent diabetes mellitus. The participating systems should be able to determine if the text is describing a substance originated within the organism or, on the contrary, it describes a process in which the substance is used for a specific purpose and thus should be identified as pharmacological substance.

4 Task 9.2: Extraction of drug-drug interactions.

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 pairs of interacting drugs, DDIExtraction 2013 also pursues the classification of each drug-drug interaction according to one of the following four types: *advice*, *effect*, *mechanism*, *int*. A detailed description of these

[3] http://www.cs.york.ac.uk/semeval-2013/task9/

[4] http://www.whocc.no/atc_ddd_index/

[5] http://www.ncbi.nlm.nih.gov/mesh

Team	Run	Rank	STRICT	EXACT	PARTIAL	TYPE	DRUG	BRAND	GROUP	DRUG_N	MAVG
LASIGE	1	6	0,656	0,781	0,808	0,69	0,741	0,581	0,712	0,171	0,577
	2	9	0,639	0,775	0,801	0,672	0,716	0,541	0,696	0,182	0,571
	3	10	0,612	0,715	0,741	0,647	0,728	0,354	0,647	0,16	0,498
NLM_LHC	1	4	0,698	0,784	0,801	0,722	0,803	0,809	0,646	0	0,57
	2	3	0,704	0,792	0,807	0,726	**0,81**	0,846	0,643	0	0,581
UMCC_DLSI	1,2,3	14,15,16	0,275	0,3049	0,367	0,334	0,297	0,313	0,257	0,124	0,311
UEM_UC3M	1	13	0,458	0,528	0,585	0,51	0,718	0,075	0,291	0,185	0,351
	2	12	0,529	0,609	0,669	0,589	0,752	0,094	0,291	**0,264**	0,38
UTurku	1	11	0,579	0,639	0,719	0,701	0,721	0,603	0,478	0,016	0,468
	2	8	0,641	0,659	0,731	0,766	0,784	0,901	0,495	0,015	0,557
	3	7	0,648	0,666	0,743	**0,777**	0,783	**0,912**	0,485	0,076	0,604
WBI	1	5	0,692	0,772	0,807	0,729	0,768	0,787	0,761	0,071	0,615
	2	2	0,708	0,831	0,855	0,741	0,786	0,803	0,757	0,134	**0,643**
	3	1	**0,715**	**0,833**	**0,856**	0,748	0,79	0,836	**0,776**	0,141	0,652

Table 3: F1 scores for task 9.1 on the whole test dataset (DDI-MedLine + DDI-DrugBank). (MAVG for macro-average). Each run is ranked by STRICT performance.

Team	Run	Rank	STRICT	EXACT	PARTIAL	TYPE	DRUG	BRAND	GROUP	DRUG_N	MAVG
LASIGE	1	8	0,771	0,834	0,855	0,799	0,817	0,571	**0,833**	0	0,563
	2	9	0,771	0,831	0,852	0,799	0,823	0,553	0,824	0	0,568
	3	11	0,682	0,744	0,764	0,713	0,757	0,314	0,756	0	0,47
NLM_LHC	1	2	0,869	0,902	**0,922**	0,902	0,909	0,907	0,766	0	0,646
	2	3	0,869	**0,903**	0,919	0,896	0,911	0,907	0,754	0	0,644
UMCC_DLSI	1,2,3	14,15,16	0,424	0,4447	0,504	0,487	0,456	0,429	0,371	0	0,351
UEM_UC3M	1	13	0,561	0,632	0,69	0,632	0,827	0,056	0,362	0,022	0,354
	2	12	0,595	0,667	0,721	0,667	0,842	0,063	0,366	**0,028**	0,37
UTurku	1	10	0,739	0,753	0,827	0,864	0,829	0,735	0,553	0	0,531
	2	6	0,785	0,795	0,863	**0,908**	0,858	0,898	0,559	0	0,581
	3	7	0,781	0,787	0,858	0,905	0,847	**0,911**	0,551	0	0,578
WBI	1	5	0,86	0,877	0,9	0,89	0,905	0,857	0,782	0	0,636
	2	4	0,868	0,894	0,914	0,897	0,909	0,865	0,794	0	0,642
	3	1	**0,878**	0,901	0,917	**0,908**	**0,912**	0,904	0,806	0	**0,656**

Table 4: F1 scores for task 9.1 on the DDI-DrugBank test data. (MAVG for macro-average). Each run is ranked by STRICT performance.

Team	Run	Rank	STRICT	EXACT	PARTIAL	TYPE	DRUG	BRAND	GROUP	DRUG_N	MAVG
LASIGE	1	4	0,567	0,74	0,772	0,605	0,678	0,667	0,612	0,183	**0,577**
	2	8	0,54	0,733	0,763	0,576	0,631	0,444	0,595	0,196	0,512
	3	6	0,557	0,693	0,723	0,596	0,702	0,667	0,56	0,171	0,554
NLM_LHC	1	5	0,559	0,688	0,702	0,575	0,717	0,429	0,548	0	0,462
	2	3	0,569	0,702	0,715	0,586	**0,726**	0,545	0,555	0	0,486
UMCC_DLSI	1,2,3	14,15,16	0,187	0,2228	0,287	0,245	0,2	0,091	0,191	0,13	0,23
UEM_UC3M	1	13	0,39	0,461	0,516	0,431	0,618	0,111	0,238	0,222	0,341
	2	11	0,479	0,564	0,628	0,529	0,665	0,182	0,233	**0,329**	0,387
UTurku	1	12	0,435	0,538	0,623	0,556	0,614	0,143	0,413	0,016	0,328
	2	10	0,502	0,528	0,604	0,628	0,703	**0,923**	0,436	0,016	0,533
	3	9	0,522	0,551	0,634	**0,656**	0,716	**0,923**	0,426	0,08	0,582
WBI	1	7	0,545	0,681	0,726	0,589	0,634	0,353	0,744	0,074	0,479
	2	2	0,576	**0,779**	**0,807**	0,612	0,673	0,444	0,729	0,14	0,534
	3	1	**0,581**	0,778	0,805	0,617	0,678	0,444	**0,753**	0,147	0,537

Table 5: F1 scores for task 9.1 on the DDI-MedLine test data. (MAVG for macro-average). Each run is ranked by STRICT performance.

types can be found in our annotation guidelines[6].

Gold standard annotations (correct, human-created annotations) of pharmacological substances are provided to participants both for training and test data. The test data for this subtask consists of 158 DrugBank documents and 33 MedLine abstracts. 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 otherwise) and its type (label *null* when the prediction value is 0), and formatted as:

IdSentence|IdDrug1|IdDrug2|prediction|type

4.1 Evaluation Metrics

Evaluation is relation-oriented and based on the standard precision, recall and F-score metrics. 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. The performance of systems to identify those pairs of drugs interacting (regardless of the type) is also evaluated. This allows us to assess the progress made with regard to the previous edition, which only dealt with the detection of DDIs.

Additionally, we are interested in assessing which drug interaction types are most difficult to detect. Thus, we calculate precision, recall and F1 for each DDI type and then their macro-average measures are provided. While micro-averaged F1 is 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 results.

Evaluating each DDI type separately allows us to assess the level of difficulty of detecting and classifying 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.

4.2 Results and Discussion

The task of extracting drug-drug interactions from biomedical texts has attracted the participation of 8 teams (see Table 2) who submitted 22 runs. Tables 6, 7 and 8 show the results for each run in alphabetic order. Due to the lack of space, the performance information is only shown in terms of F1 score. The reader can find the full ranking information on the SemEval-2013 Task 9 website[7].

Most of the participating systems were built on support vector machines. In general, approaches based on non-linear kernels methods achieved better results than linear SVMs. As in the previous edition of DDIExtraction, most systems have used primarily syntactic information. However, semantic information has been poorly used.

The best results were submitted by the team from FBK-irst. They applied a novel hybrid kernel based RE approach described in Chowdhury (2013a). They also exploited the scope of negations and semantic roles for negative instance filtering as proposed in (Chowdhury and Lavelli, 2013b) and (Chowdhury and Lavelli, 2012). The second best results were obtained by the WBI team from the Humboldt University of Berlin. Its system combines several kernel methods (APG (Airola et al., 2008) and Shallow Linguistic Kernel (SL) (Giuliano et al., 2006) among others), the Turku Event Extraction system (TEES) (Björne et al., 2011)[8] and the Moara system (Neves et al., 2009). These two teams were also the top two ranked teams in DDIExtraction 2011. For a more detailed description, the reader is encouraged to read the papers of the participants in the proceedings book.

While the DDIExtraction 2011 shared task concentrated efforts on the detection of DDIs, this new DDIExtraction 2013 task involved not only the detection of DDIs, but also their classification. Although the results of DDIExtraction 2011 are not directly comparable with the ones reported in DDIExtraction 2013 due to the use of different training and test datasets in each edition, it should be noted that there has been a significant improvement in the de-

[6] http://www.cs.york.ac.uk/semeval-2013/task9/

[7] http://www.cs.york.ac.uk/semeval-2013/task9/

[8] http://jbjorne.github.io/TEES/

Team	Run	Rank	CLA	DEC	MEC	EFF	ADV	INT	MAVG
FBK-irst	1	3	0.638	**0.8**	**0.679**	**0.662**	**0.692**	0.363	0.602
	2	1	**0.651**	**0.8**	**0.679**	0.628	**0.692**	**0.547**	**0.648**
	3	2	0.648	**0.8**	0.627	**0.662**	**0.692**	**0.547**	0.644
NIL_UCM	1	12	0.517	0.588	0.515	0.489	0.613	0.427	0.535
	2	10	0.548	0.656	0.531	0.556	0.61	0.393	0.526
SCAI	1	14	0.46	0.69	0.446	0.459	0.562	0.02	0.423
	2	16	0.452	0.683	0.441	0.44	0.559	0.021	0.448
	3	15	0.458	0.704	0.45	0.462	0.54	0.02	0.411
UC3M	1	11	0.529	0.676	0.48	0.547	0.575	0.5	0.534
	2	21	0.294	0.537	0.268	0.286	0.325	0.402	0.335
UCOLORADO_SOM	1	22	0.214	0.492	0.109	0.25	0.219	0.097	0.215
	2	20	0.334	0.504	0.361	0.311	0.381	0.333	0.407
	3	19	0.336	0.491	0.335	0.313	0.42	0.329	0.38
UTurku	1	9	0.581	0.684	0.578	0.585	0.606	0.503	0.572
	2	7	0.594	0.696	0.582	0.6	0.63	0.507	0.587
	3	8	0.582	0.699	0.569	0.593	0.608	0.511	0.577
UWM-TRIADS	1	17	0.449	0.581	0.413	0.446	0.502	0.397	0.451
	2	13	0.47	0.599	0.446	0.449	0.532	0.421	0.472
	3	18	0.432	0.564	0.442	0.383	0.537	0.292	0.444
WBI	1	6	0.599	0.736	0.602	0.604	0.618	0.516	0.588
	2	5	0.601	0.745	0.616	0.595	0.637	0.49	0.588
	3	4	0.609	0.759	0.618	0.61	0.632	0.51	0.597

Table 6: F1 scores for Task 9.2 on the whole test dataset (DDI-MedLine + DDI-DrugBank). DEC for Detection, CLA for detection and classification, MEC for *mechanism* type, EFF for *effect* type, ADV for *advice* type, INT for *int* type and MAVG for macro-average. Each run is ranked by CLA performance.

Team	Run	Rank	CLA	DEC	MEC	EFF	ADV	INT	MAVG
FBK-irst	1	3	0.663	**0.827**	**0.705**	**0.699**	**0.705**	0.376	0.624
	2	1	**0.676**	**0.827**	**0.705**	0.664	**0.705**	**0.545**	**0.672**
	3	2	0.673	**0.827**	0.655	**0.699**	**0.705**	**0.545**	0.667
NIL_UCM	1	12	0.54	0.615	0.527	0.525	0.625	0.444	0.565
	2	10	0.573	0.68	0.552	0.597	0.619	0.408	0.55
SCAI	1	15	0.464	0.711	0.449	0.459	0.57	0.021	0.461
	2	16	0.463	0.71	0.445	0.458	0.569	0.021	0.46
	3	14	0.473	0.734	0.468	0.482	0.551	0.021	0.439
UC3M	1	11	0.555	0.703	0.493	0.593	0.59	0.51	0.561
	2	21	0.306	0.549	0.274	0.302	0.334	0.426	0.352
UCOLORADO_SOM	1	22	0.218	0.508	0.115	0.251	0.24	0.098	0.228
	2	20	0.341	0.518	0.373	0.313	0.398	0.344	0.425
	3	19	0.349	0.511	0.353	0.324	0.429	0.327	0.394
UTurku	1	8	0.608	0.712	0.6	0.63	0.617	0.522	0.6
	2	7	0.62	0.724	0.605	0.644	0.638	0.522	0.614
	3	9	0.608	0.726	0.591	0.635	0.617	0.522	0.601
UWM-TRIADS	1	17	0.462	0.596	0.43	0.459	0.509	0.405	0.463
	2	13	0.485	0.616	0.467	0.466	0.536	0.425	0.486
	3	18	0.445	0.573	0.469	0.39	0.544	0.29	0.46
WBI	1	6	0.624	0.762	0.621	0.645	0.634	0.52	0.61
	2	5	0.627	0.775	0.636	0.636	0.652	0.5	0.611
	3	4	0.632	0.783	0.629	0.652	0.65	0.513	0.617

Table 7: F1 scores for task 9.2 on the DDI-DrugBank test dataset. Each run is ranked by CLA performance.

Team	Run	Rank	CLA	DEC	MEC	EFF	ADV	INT	MAVG
FBK-irst	1	4	0.387	**0.53**	0.383	0.436	0.286	0.211	0.406
	2	3	0.398	**0.53**	0.383	0.407	0.286	**0.571**	0.436
	3	2	0.398	**0.53**	0.339	0.436	0.286	**0.571**	**0.44**
NIL_UCM	1	20	0.19	0.206	0.286	0.186	0	0	0.121
	2	19	0.219	0.336	0.143	0.271	0	0	0.11
SCAI	1	1	**0.42**	0.462	0.412	**0.458**	0.2	0	0.269
	2	8	0.323	0.369	0.389	0.333	0	0	0.182
	3	6	0.341	0.474	0.31	0.379	0.222	0	0.229
UC3M	1	15	0.274	0.406	0.333	0.267	0	0.364	0.268
	2	22	0.186	0.421	0.222	0.171	0.143	0	0.149
UCOLORADO_SOM	1	21	0.188	0.37	0.042	0.241	0	0	0.073
	2	14	0.275	0.394	0.258	0.302	0.138	0	0.177
	3	17	0.244	0.356	0.194	0.255	0.222	0.4	0.272
UTurku	1	18	0.242	0.339	0.258	0.256	0.2	0	0.18
	2	16	0.262	0.344	0.214	0.278	0.364	0	0.224
	3	13	0.286	0.376	0.286	0.289	0.333	0	0.232
UWM-TRIADS	1	10	0.312	0.419	0.233	0.36	0.267	0	0.219
	2	9	0.319	0.436	0.233	0.34	**0.421**	0.333	0.345
	3	11	0.306	0.479	0.247	0.326	0.381	0.333	0.33
WBI	1	7	0.336	0.456	0.368	0.344	0.154	0.4	0.334
	2	12	0.304	0.406	0.343	0.318	0.167	0	0.209
	3	5	0.365	0.503	**0.476**	0.347	0.143	0.4	0.353

Table 8: F1 scores for task 9.2 on the DDI-MedLine test dataset. Each run is ranked by CLA performance.

tection of DDIs: F1 has a remarkable increase from 65.74% (the best F1-score in DDIExtraction 2011) to 80% (see *DEC* column of Table 6). The increase of the size of the corpus made for DDIExtraction 2013 and of the quality of their annotations may have contributed significantly to this improvement.

However, the results for the detection and classification for DDIs did not exceed an F1 of 65.1%. Table 6 suggests that some type of DDIs are more difficult to classify than others. The best F1 ranges from 69.2% for *advice* to 54.7% for *int*. One possible explanation for this could be that recommendations or advice regarding a drug interaction are typically described by very similar text patterns such as *DRUG should not be used in combination with DRUG* or *Caution should be observed when DRUG is administered with DRUG*.

Regarding results for the *int* relationship, it should be noted that the proportion of instances of this relationship (5.6%) in the DDI corpus is much smaller than those of the rest of the relations (41.1% for *effect*, 32.3% for *mechanism* and 20.9% for *advice*).

As stated earlier, one of the differences from the previous edition is that the corpus developed for DDIExtraction 2013 is made up of texts from two different sources: MedLine and the DrugBank database. Thus, the different approaches can be evaluated on two different styles of biomedical texts. While MedLine abstracts are usually written in extremely scientific language, texts from DrugBank are written in a less technical form of the language (similar to the language used in package inserts). Indeed, this may be the reason why the results on the DDI-DrugBank dataset are much better than those obtained on the DDI-MedLine dataset (see Tables 7 and 8).

5 Conclusions

The DDIExtraction 2011 task concentrated efforts on the novel aspects of the DDI extraction task, the drug recognition was assumed and the annotations for drugs were provided to the participants. This new DDIExtraction 2013 task pursues the detection and classification of drug interactions as well as the recognition and classification of pharmacological substances. The task attracted broad interest from the community. A total of 14 teams from 7 different countries participated, submitted a total of 38 runs, exceeding the participation of DDIExtraction 2011 (10 teams). The participating systems demonstrated substantial progress at the established DDI extraction task on DrugBank texts and showed that their methods also obtain good results for MedLine abstracts.

The results that the participating systems have reported show successful approaches to this difficult task, and the advantages of non-linear kernel-based methods over linear SVMs for extraction of DDIs. In the named entity task, the participating systems perform well in recognizing generic drugs, brand drugs and groups of drugs, but they fail in recognizing active substances not approved for human use. Although the results are positive, there is still much room to improve in both subtasks. We have accomplished our goal of providing a framework and a benchmark data set to allow for comparisons of methods for the recognition of pharmacological substances and detection and classification of drug-drug interactions from biomedical texts.

We would like that our test dataset can still serve as the basis for fair and stable evaluation after the task. Thus, we have decided that the full gold annotations for the test data are not available for the moment. We plan to make available a web service where researchers can test their methods on the test dataset and compare their results with the DDIExtraction 2013 task participants.

Acknowledgments

This research work has been supported by the Regional Government of Madrid under the Research Network MA2VICMR (S2009/TIC-1542), by the Spanish Ministry of Education under the project MULTIMEDICA (TIN2010-20644-C03-01). Additionally, we would like to thank all participants for their efforts and to congratulate them to their interesting work.

References

A. Airola, S. Pyysalo, J. Bjorne, T. Pahikkala, F. Ginter, and T. Salakoski. 2008. All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC bioinformatics*, 9(Suppl 11):S2.

JK. Aronson. 2007. Communicating information about drug interactions. *British Journal of Clinical Pharmacology*, 63(6):637–639, June.

K. Baxter and I.H. Stockely. 2010. *Stockley's drug interactions.8th ed.* London:Pharmaceutical Press.

J. Björne, J. Heimonen, F. Ginter, A. Airola, T. Pahikkala, and T. Salakoski. 2011. Extracting contextualized complex biological events with graph-based feature sets. *Computational Intelligence*, 27(4):541–557.

J. Björne, S. Kaewphan, and T. Salakoski. 2013. UTurku: Drug Named Entity Detection and Drug-drug Interaction Extraction Using SVM Classification and Domain Knowledge. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

T. Bobić, J. Fluck, and M. Hofmann-Apitius. 2013. SCAI: Extracting drug-drug interactions using a rich feature vector. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

A. Bokharaeian, B.and Díaz. 2013. NIL_UCM: Extracting Drug-Drug interactions from text through combination of sequence and tree kernels. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

D. Boring. 1997. The development and adoption of nonproprietary, established, and proprietary names for pharmaceuticals. *Drug information journal*, 31(3):621–634.

N. Chinchor and P. Robinson. 1997. Muc-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding*.

N. Chinchor and B. Sundheim. 1993. Muc-5 evaluation metrics. In *Proceedings of the 5th conference on Message understanding*, pages 69–78. Association for Computational Linguistics.

MFM. Chowdhury and A. Lavelli. 2012. Impact of less skewed distributions on efficiency and effectiveness of biomedical relation extraction. In *Proceedings of COLING 2012*.

MFM. Chowdhury and A. Lavelli. 2013b. Exploiting the scope of negations and heterogeneous features for relation extraction: Case study drug-drug interaction extraction. In *Proceedings of NAACL 2013*.

M.F.M. Chowdhury and A. Lavelli. 2013c. FBK-irst : A Multi-Phase Kernel Based Approach for Drug-Drug Interaction Detection and Classification that Exploits Linguistic Information. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

MFM. Chowdhury. 2013a. *Improving the Effectiveness of Information Extraction from Biomedical Text*. Ph.d. dissertation, University of Trento.

A. Collazo, A. Ceballo, D Puig, Y. Gutiérrez, J. Abreu, J Pérez, A. Fernández-Orquín, A. Montoyo, R. Muñoz, and F. Camara. 2013. UMCC_DLSI-(DDI): Semantic and Lexical features for detection and classification Drugs in biomedical texts. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

C. Giuliano, A. Lavelli, and L. Romano. 2006. Exploiting shallow linguistic information for relation extraction from biomedical literature. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, pages 401–408.

T. Grego, F. Pinto, and F.M. Couto. 2013. LASIGE: using Conditional Random Fields and ChEBI ontology. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

N.D. Hailu, L.E. Hunter, and K.B. Cohen. 2013. UColorado_SOM: Extraction of Drug-Drug Interactions from Biomedical Text using Knowledge-rich and Knowledge-poor Features. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

ML. Neves, JM. Carazo, and A. Pascual-Montano. 2009. Extraction of biomedical events using case-based reasoning. In *Proceedings of the Workshop on BioNLP: Shared Task*, pages 68–76. Association for Computational Linguistics.

S. Pyysalo, A. Airola, J. Heimonen, J. Bjorne, F. Ginter, and T. Salakoski. 2008. Comparative analysis of five protein-protein interaction corpora. *BMC bioinformatics*, 9(Suppl 3):S6.

M. Rastegar-Mojarad, R. D. Boyce, and R. Prasad. 2013. UWM-TRIADS: Classifying Drug-Drug Interactions with Two-Stage SVM and Post-Processing. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

T. Rocktäschel, M. Weidlich, and U. Leser. 2012. Chemspot: a hybrid system for chemical named entity recognition. *Bioinformatics*, 28(12):1633–1640.

T. Rocktäschel, T. Huber, M. Weidlich, and U. Leser. 2013. WBI-NER: The impact of domain-specific features on the performance of identifying and classifying mentions of drugs. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

D. Sanchez-Cisneros and F. Aparicio. 2013. UEM-UC3M: An Ontology-based named entity recognition system for biomedical texts. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

D. Sanchez-Cisneros. 2013. UC3M: A kernel-based approach for identify and classify DDIs in biomedical

texts. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

I. Segura-Bedmar, P. Martínez, and C. de Pablo-Sánchez. 2011a. Using a shallow linguistic kernel for drug-drug interaction extraction. *Journal of Biomedical Informatics*, 44(5):789 – 804.

I. Segura-Bedmar, P. Martınez, and D. Sánchez-Cisneros. 2011b. The 1st ddiextraction-2011 challenge task: Extraction of drug-drug interactions from biomedical texts. In *Proceedings of DDIExtraction-2011 challenge task*, pages 1–9.

P. Thomas, M. Neves, T. Rocktäschel, and U. Leser. 2013. WBI-DDI: Drug-Drug Interaction Extraction using Majority Voting. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

E.F. Tjong Kim Sang and F. De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.

D.S. Wishart, C. Knox, A.C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey. 2006. Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, 34(suppl 1):D668–D672.