Overview of the TAC 2019 Track on Drug-Drug Interaction Extraction from Drug Labels

Travis Reed Goodwin
National Institutes of Health

Overview of the TAC 2019 Track on Drug-Drug Interaction Extraction from Drug Labels

Travis R. Goodwin[1], Dina Demner-Fushman[1], Kin Wah Fung[1], and Phong Do[2]

[1] U.S. National Library of Medicine
nlmlhclhcques@mail.nih.gov
https://www.nlm.nih.gov
[2] Office of Health Informatics
U.S. Food and Drug Administration

Abstract. This paper describes the Drug-Drug Interaction Extraction from Drug Labels Track, part of the 2019 Text Analysis Conference (TAC). Participants were provided with an annotated set of interactions-related sections of drug labels and challenged with: (1) extracting mentions of the precipitants and effects of drug-drug interactions at the sentence level; (2) identifying (typed) relations between interacting substances; (3) normalizing mentions and relations to several standard terminologies; and (4) determining the unique set of drug-drug interactions across all provided sections of a drug label. Four teams submitted at least one run, with 10 submissions in total.

1 Objective

The purpose of the Drug-Drug Interaction Extraction from Drug Labels (DDI) track in TAC 2019 is to evaluate various natural language processing (NLP) approaches based on their information extraction (IE) performance when identifying drug-drug interactions from Structured Product Labeling (SPL) documents. SPL is a document markup standard approved by Health Level Seven (HL7) and adopted by the FDA as a mechanism for exchanging product and facility information about drugs. In this paper, we provide background on the DDI track, describe the dataset and SPL annotation effort, provide an overview of each of the four tasks we evaluated, give an overview of participating teams, and present the results of the 2019 TAC DDI evaluation.

2 Background

The U.S. Food and Drug Administration (FDA) is responsible for protecting public health by assuring safety, efficacy, and security of all FDA-regulated products, including human and veterinary drugs, prescription and over-the-counter pharmaceutical drugs, vaccines, biopharmaceuticals, blood transfusions, and biological products, among others. FDA and the National Library of Medicine (NLM) have been working together on transforming the content of Structured

Product Labeling (SPL) documents for prescription drugs into discrete, coded, computer-readable data that will be made available to the public in individual SPL index documents. Transforming the narrative text to structured information encoded in national standard terminologies is a prerequisite to the effective deployment of drug safety information. Being able to electronically access labeling information and to search and sort that information is an important step toward creation of a fully automated health information exchange system. TAC 2017 addressed one of the important drug safety issues: automated extraction of adverse drug reactions reported in SPLs [1]. An equally important and complex task is automated extraction of drug-drug interaction (DDI) information. Drug-drug interactions can lead to a variety of adverse events, and it has been suggested that preventable adverse events are the eighth leading cause of death in the United States [2].

Structuring drug safety information is a task in which natural language processing (NLP) systems can provide a great benefit to the FDA and medical community in general. The purpose of this TAC track is to test various NLP approaches for their information extraction (IE) performance on drug-drug interactions in SPLs. While the ultimate goal is for NLP systems to extract and code to controlled terminologies the distinct interactions from the drug labels (the standard structured representation for drug interactions), this track also evaluates and provides data for several intermediate tasks, such as extracting entities (e.g., substances and effects) and relations, as well as normalizing the extracted terms and relations to the FDA substance registration system Unique Ingredient Identifiers (UNII), Medication Reference Terminology (MED-RT), SNOMED CT, and NCI Thesaurus pharmacokinetic effects. The results of this track will inform future FDA efforts at automating important safety processes.

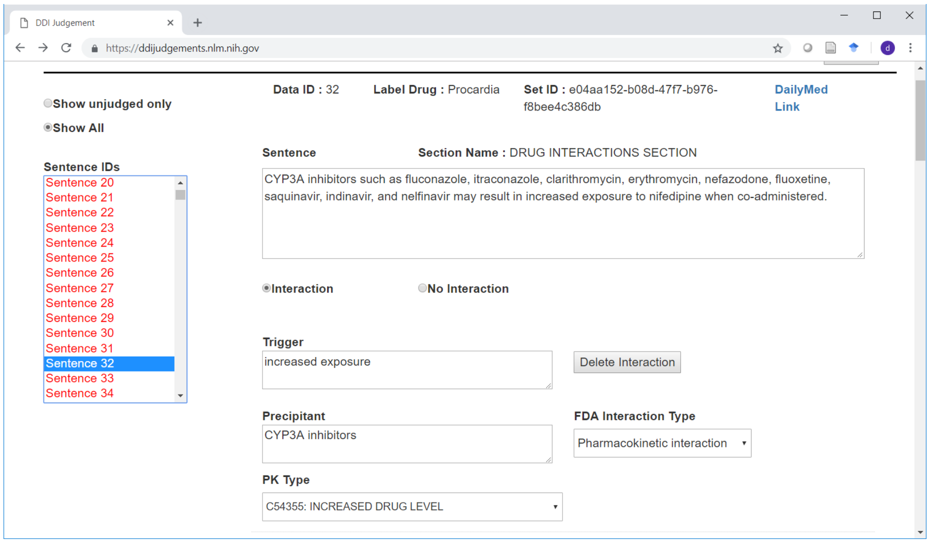
2.1 Related Work

Earlier work on DDI extraction from SPLs provided some potentially useful training data [3, 4], although none of the previous annotations exactly match the FDA requirements for structuring DDIs for the SPL index files. Data sets prepared for the TAC 2018 DDI track and the approaches explored by the participating teams are, of course, addressing the tasks at hand. In addition to extraction of DDIs from SPLs, two information extraction areas are closely related to the DDI TAC 2019 track: extraction of other information from SPLs and extraction of DDI from other types of text, e.g., literature and social media. DDI Extraction Challenges 2011 and 2013 focused on extracting DDI information from the literature [5]. These challenges and datasets facilitated a growing body of research, with the latest recursive neural network model that implements a tree-LSTM architecture achieving 83.8% F1-score for DDI detection and 73.5% F1-score for interaction type classification [6]. Other types of information that need to be extracted from SPLs include adverse drug reactions [1], indications [7], use in special populations [8], and several others, e.g., pharmacogenomics biomarkers or the drug's mechanism of action, that have not been explored yet.

3 Data

The TAC 2019 DDI track dataset consists of 406 human-annotated Structured Product Labels, in which most or some of the following sections are annotated with drug-drug interactions: *Boxed Warning*, *Clinical Pharmacology*, *Contraindications*, *Dosage and Administration*, *Drug and/or laboratory test interaction*, *Drug Interactions*, *Precautions*, *Warnings and Precautions* and *Warnings*. The dataset was divided into 325 fully or partially annotated SPLs provided to participants for training, and a set of 81 SPLs with annotations withheld from participants used to evaluate submissions.

All training and test set files follow the TAC-specific XML format that exactly follows the evaluation schema and annotation requirements. All interactions are annotated with respect to the Labeled Drug, i.e., the drug for which the SPL was published. Some of the annotations in the training set were generated semi-automatically and might be missing some interactions. FDA experts and NLM staff and volunteers manually corrected the automatically extracted entities and relations using the interface in Fig. 1.



○Show unjudged only

●Show All

Data ID : 32 **Label Drug :** Procardia **Set ID :** e04aa152-b08d-47f7-b976-f8bee4c386db

DailyMed Link

Sentence IDs

Sentence 20
Sentence 21
Sentence 22
Sentence 23
Sentence 24
Sentence 25
Sentence 26
Sentence 27
Sentence 28
Sentence 29
Sentence 30
Sentence 31
Sentence 32
Sentence 33
Sentence 34

Sentence **Section Name :** DRUG INTERACTIONS SECTION

CYP3A inhibitors such as fluconazole, itraconazole, clarithromycin, erythromycin, nefazodone, fluoxetine, saquinavir, indinavir, and nelfinavir may result in increased exposure to nifedipine when co-administered.

●Interaction ○No Interaction

Trigger
increased exposure

Delete Interaction

Precipitant
CYP3A inhibitors

FDA Interaction Type
Pharmacokinetic interaction

PK Type
C54355: INCREASED DRUG LEVEL

Fig. 1. DDI annotation interface. The online interface for registered users to annotate label sentences assigned to them. The full SPL can be reached using the DailyMed link in the upper right corner.

The evaluation set was fully manually annotated by FDA and NLM using the guidelines finalized before annotation[3].

[3] `https://bionlp.nlm.nih.gov/tac2019druginteractions/DDIvalidationGuidelines.docx`

Box 1: Entity Annotations

The following entities are annotated in the gold standard:

Precipitant A substance interacting with the Labeled Drug could be another drug, a drug class or a non-drug substance (e.g., *alcohol*, *grapefruit juice*.)

Trigger A word or phrase indicating an interaction event.

SpecificInteraction Results of interactions, e.g., *severe hyperkalemia*.

Box 2: Interaction Annotations

The following interaction relations connect the above entities in an Interaction. Each relation is limited to a specific subset of entity types.

Pharmacokinetic interactions (PK) between the Labeled Drug and the precipitant are indicated by Triggers, e.g., *reducing diuretic absorption*, and other phrases indicating increases / decreases in function measurements.

Pharmacodynamic (Specific) interactions between the Labeled Drug and the precipitant are indicated by Triggers, e.g., *potentiate* or *increased risks* and result in SpecificInteraction.

Unspecified interactions are indicated by Triggers, e.g., *avoid use*.

3.1 Annotations

The entities and sentence-level interactions in annotated SPLs are indicated in Boxes 1 and 2. Box 3 provides information about the controlled vocabularies and terminologies used for normalizing entities and interactions. Note: unlike previous years, annotators were instructed to provide multiple mappings for effects if there were multiple valid mappings in the source vocabulary.

The ultimate goal of the task is to know which interactions are present in the SPL documents such that the interactions may be linked to structured knowledge sources. An interaction mentioned several times should not necessarily carry more weight than an interaction mentioned once. Consequently, to test the systems on finding distinct interactions, the gold standard contains a list of unique normalized interactions aggregated at the document level.

4 Tasks

The track contained four specific tasks, each one potentially building upon the previous tasks:

Task 1 Extract Mentions of Interacting Drugs/Substances, and specific interactions at sentence level. This is similar to many NLP named entity

Box 3: Normalization

The entities and interactions are mapped as follows:

- The interacting substances are mapped to UNIIs.
- Drug classes are mapped to MED-RT NUIs.
- The effect of the interaction is mapped to a SNOMED CT CUI, if it is a medical condition.
- Pharmacokinetic effects are mapped to National Cancer Institute Thesaurus codes.

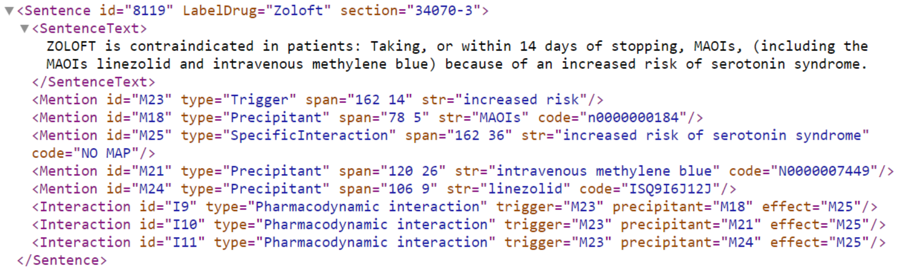
recognition (NER) evaluations. Note: mentions of interaction triggers are not evaluated in the 2019 DDI track.

Task 2 Identify interactions at sentence level, including: the interacting drugs, the specific interaction types: pharmacokinetic, pharmacodynamic or unspecified, and the outcomes of pharmacokinetic and pharmacodynamic interactions. This is similar to many NLP relation identification evaluations.

Task 3 Normalization task. The interacting substance should be normalized to UNII, and the drug classes to MED-RT*. The consequence of the interaction should be normalized to SNOMED CT if it is a medical condition. Pharmacokinetic effects are normalized to National Cancer Institute Thesaurus codes. Note: Drug classes are mapped to MED-RT rather than NDF-RT for the 2019 DDI track.

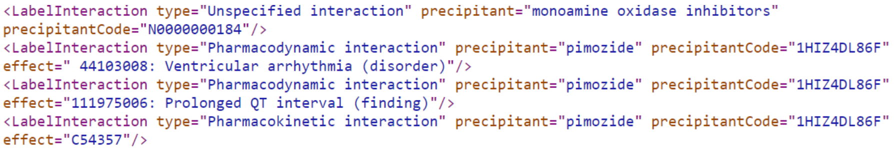
Task 4 Generate a global list of distinct interactions in normalized form for each label.

Tasks 1, 2 and 3 correspond to traditional NLP information extraction (IE) and entity linking tasks, while Task 4 involves document-level aggregation. While the tasks were designed to build on each other, participation was optional on a per-task basis. See Fig. 2 and Fig. 3 for examples of the sentence- and document-level annotations expected from the participating systems.



```
▼<Sentence id="8119" LabelDrug="Zoloft" section="34070-3">
  ▼<SentenceText>
     ZOLOFT is contraindicated in patients: Taking, or within 14 days of stopping, MAOIs, (including the
     MAOIs linezolid and intravenous methylene blue) because of an increased risk of serotonin syndrome.
   </SentenceText>
   <Mention id="M23" type="Trigger" span="162 14" str="increased risk"/>
   <Mention id="M18" type="Precipitant" span="78 5" str="MAOIs" code="n0000000184"/>
   <Mention id="M25" type="SpecificInteraction" span="162 36" str="increased risk of serotonin syndrome"
   code="NO MAP"/>
   <Mention id="M21" type="Precipitant" span="120 26" str="intravenous methylene blue" code="N0000007449"/>
   <Mention id="M24" type="Precipitant" span="106 9" str="linezolid" code="ISQ9I6J12J"/>
   <Interaction id="I9" type="Pharmacodynamic interaction" trigger="M23" precipitant="M18" effect="M25"/>
   <Interaction id="I10" type="Pharmacodynamic interaction" trigger="M23" precipitant="M21" effect="M25"/>
   <Interaction id="I11" type="Pharmacodynamic interaction" trigger="M23" precipitant="M24" effect="M25"/>
 </Sentence>
```

Fig. 2. Sentence-level annotations of pharmacodynamics interactions between Zoloft and Monoamine oxidase inhibitors (MAOIs). Three precipitants cause the same effect indicated by the same trigger, which results in three annotated interactions.



```
<LabelInteraction type="Unspecified interaction" precipitant="monoamine oxidase inhibitors"
precipitantCode="N0000000184"/>
<LabelInteraction type="Pharmacodynamic interaction" precipitant="pimozide" precipitantCode="1HIZ4DL86F"
effect=" 44103008: Ventricular arrhythmia (disorder)"/>
<LabelInteraction type="Pharmacodynamic interaction" precipitant="pimozide" precipitantCode="1HIZ4DL86F"
effect="111975006: Prolonged QT interval (finding)"/>
<LabelInteraction type="Pharmacokinetic interaction" precipitant="pimozide" precipitantCode="1HIZ4DL86F"
effect="C54357"/>
```

Fig. 3. Document-level annotations of all types of interactions between Zoloft, MAOIs and pimozide.

5 Evaluation

Submitted systems were evaluated using the following task-specific measures:

Task 1 Precision/Recall/F_1-measure on annotated entities (substances and effects) using the surface form of mentions (i.e., not offset dependent). Both mentions with type and without type were evaluated. The primary evaluation metric was micro-averaged F_1 across the exact matched entity-level annotations with type.

Task 2 Precision/Recall/F_1-measure on relations. Both the full relation (all elements of interaction, i.e., the precipitant, the effect and the interaction type) and the presence of relations were evaluated, both with and without type. The primary evaluation metric was micro-averaged F_1 across full relations with type.

Task 3 Precision/Recall/F_1-measure on linking entities to the specific terminologies. The primary evaluation metric was F_1 macro-averaged across SPLs.

Task 4 SPL-level Precision/Recall/F_1-measure on unique normalized interactions. The primary evaluation metric was F_1 macro-averaged across labels.

6 Participants

Four teams participated in the 2019 edition of the DDI track:

IBMResearch *IBM Research*. The team participated in Tasks 1 and 2 using a pre-trained language model approach for entity extraction and interaction identification. This team also experimented with dependency-parse-based post-processing.

INK_BC *Shandong University of Finance and Economics* (Chinese: 山东财经大学). The team participated in Tasks 1 and 2 using a hybrid approach combining context and n-gram models.

UTDHLTRI *The Human Language Technology Research Institute (HLTRI) at the University of Texas at Dallas (UTD)*. The only team to participate in all four tasks, the team used a combination of Deep Bidirectional Transformers for Language Understanding (BERT)[10] and Conditional Random Fields (CRFs)[11] for task 1, BERT and multi-task relation extraction for tasks 2 and 4, and string-based pattern matching for task 3.

SRCB *Ricoh Software Research Center (Beijing)* (Chineese: 理光软件研究所(北京)有限公司). The team participated in tasks 1, 2, and 3. They approached task 1 using a BERT-based model with additional universal transformer layers and automatic data augmentation and relative position attention. Task 2 used BERT with interaction-type-based context and syntactic features. Task 3 was approached using Apache Solr[4] and multiple string kernels.

7 Results

The results for all runs are shown in Tables 1 − 4. Runs are sorted by primary metric (micro- or macro-average F_1 score) in descending order. Run descriptions are provided in Appendix A.

Table 1. Task 1 (Named Entity Recognition) results sorted by micro-average F_1 score. The median score was 48.98 % F_1.

Team	**Run**	**Precision**	**Recall**	**F_1-score**
		(%)	(%)	(%)
IBMResearch	1	**73.40**	58.94	**65.38**
IBMResearch	3	72.98	58.89	65.18
IBMResearch	2	73.02	57.95	64.62
SRCB	1	70.93	56.52	62.91
SRCB	3	72.46	55.52	62.87
SRCB	2	71.33	55.81	62.62
UTDHLTRI	3	24.70	60.30	35.04
UTDHLTRI	1	16.79	**67.82**	26.92
UTDHLTRI	2	16.79	**67.82**	26.92
INK_BC	1	18.15	28.73	22.25

8 Discussion

Task 4 was clearly the most challenging (attempted only by one team with the best F_1 of 17.6% compared to $F_1 \geq 49\%$ for tasks 1 and 2). This is likely due to the fact that many interactions are repeated in several sections. An optimistic view would be to assume that the most important and severe distinct interactions were captured because these are usually repeated in all annotated sections.

[4] `https://lucene.apache.org/solr/`

Table 2. Task 2 (Interaction Extraction) results sorted by micro-average F_1 score. The median score was 37.13 % F_1.

Team	**Run**	**Precision** (%)	**Recall** (%)	**F_1-score** (%)
IBMResearch	1	58.29	42.31	**49.03**
IBMResearch	3	**59.08**	40.98	48.39
IBMResearch	2	57.83	41.51	48.33
SRCB	2	54.70	40.84	46.77
SRCB	1	53.84	41.32	46.76
SRCB	3	53.84	41.32	46.76
UTDHLTRI	2	19.76	**45.09**	27.48
UTDHLTRI	3	19.76	**45.09**	27.48
UTDHLTRI	1	19.93	44.34	27.50
INK_BC	1	3.618	4.511	4.016

Table 3. Task 3 (Normalization) results sorted by macro-average F_1 score. Sentence-level scores were the primary evaluation metric in 2019, while label-level scores reflect the evaluation metric used in 2018. The median sentence-level score was 45.53 % F_1.

		Sentence-level			**Label-level**		
Team	**Run**	**Precision** (%)	**Recall** (%)	**F_1-score** (%)	**Precision** (%)	**Recall** (%)	**F_1-score** (%)
SRCB	3	**70.88**	58.49	**62.39**	**76.13**	66.62	**69.35**
SRCB	1	67.55	59.37	61.43	73.39	**67.86**	69.03
SRCB	2	65.78	56.49	59.43	71.67	66.87	67.65
UTDHLTRI	3	21.57	54.48	28.66	37.99	62.07	43.81
UTDHLTRI	1	15.20	**62.84**	22.53	27.73	66.43	36.06
UTDHLTRI	2	15.20	**62.84**	22.53	27.73	66.43	36.06

Table 4. Task 4 (Normalized Label-level Interaction Extraction) results sorted by macro-average F_1 score.

Team	**Run**	**Precision** (%)	**Recall** (%)	**F_1-score** (%)
UTDHLTRI	2	**13.41**	**31.88**	**17.56**
UTDHLTRI	3	**13.41**	**31.88**	**17.56**
UTDHLTRI	1	13.55	29.69	17.32

The results on Task 1, although the highest for this evaluation, indicate that this new task is challenging, even compared to the same DDI extraction from the literature, and needs more attention. When compared to the previous results (after adjusting for differences in the evaluation criteria), we observed a 12.36% (relative; 7.19% absolute) increase in Task 1 and an 8.80% (relative; 3.97% absolute) increase in Task 2 comparing the top performing 2018 and 2019 systems. For Task 3, we see a 236.47% (relative; 48.74% absolute) increase and for Task 4 we see a 30% (relative; 4.09% absolute) increase.[5]

Table 5. Results for Tasks 1-3 when evaluating only on sentences with (gold-standard) interactions where P indicates Precision and R indicates Recall.

Team	**Run**	**P** (%)	**R** (%)	***F*₁** (%)
UTDHLTRI	1	74.1	**67.8**	**70.8**
UTDHLTRI	2	74.1	**67.8**	**70.8**
IBMResearch	1	**84.5**	58.9	69.4
IBMResearch	3	84.2	58.9	69.3
IBMResearch	2	84.2	58.0	68.6
UTDHLTRI	3	75.6	60.3	67.1
SRCB	1	81.1	56.5	66.6
SRCB	2	81.3	55.8	66.2
SRCB	3	81.6	55.5	66.1
INK_BC	1	39.0	28.7	33.1

(a) Task 1

Team	**Run**	**P** (%)	**R** (%)	***F*₁** (%)
IBMResearch	1	61.3	42.3	**50.1**
IBMResearch	3	**62.1**	41.0	49.4
IBMResearch	2	60.9	41.5	49.4
UTDHLTRI	2	54.0	**45.1**	49.2
UTDHLTRI	3	54.0	**45.1**	49.2
UTDHLTRI	1	54.2	44.3	48.8
SRCB	1	59.1	41.3	48.6
SRCB	3	59.1	41.3	48.6
SRCB	2	60.0	40.8	48.6
INK_BC	1	7.7	4.5	5.7

(b) Task 2

Team	**Run**	**P** (%)	**R** (%)	***F*₁** (%)
SRCB	1	81.0	59.4	**67.4**
UTDHLTRI	1	73.5	**62.8**	67.1
UTDHLTRI	2	73.5	**62.8**	67.1
SRCB	3	**81.4**	58.5	66.9
SRCB	2	77.3	56.5	64.2
UTDHLTRI	3	72.3	54.5	61.4

(c) Task 3

When examining errors in submissions, we noticed that many errors resulted from teams annotating entities and interactions in sentences with no drug-drug interactions. Table 5 shows the performance of submitted systems when evaluated only on sentences containing gold-standard interactions. We can see a 8.289 % (relative) increase in F_1 and a 15.12 % (relative) increase in Precision

[5] Relative increases are computed as the percent change from the top 2018 score to the top 2019 score using the top score on the first test set from 2018 as the reference.

for Task 1. For Task 2, we observe a small (relative) increase in F_1 by 2.133 %, and for Task 3 we observe an 8.104 % (relative) increase in F_1 with a 14.86 % (relative) increase in Precision. The improvement on Tasks 1 and 3 with the minor improvement on Task 2 suggests that teams may benefit from post-processing to remove entities from sentences without any interactions. Moreover, the increase in Precision suggests that it may be worthwhile to include a sentence classification task indicating whether each sentence includes a drug-drug interaction in the future.

9 Conclusion

The goal of the TAC Drug-Drug Interaction Extraction from Drug Labels Track was to evaluate and draw attention to the important problem of identifying the drug interactions described in SPLs. Four teams submitted a total of ten runs across the four tasks. The results clearly indicate that the ultimate goal of producing index files coded to multiple terminologies fully automatically is unattainable at this time. The results achieved by half of the teams, however, show that automated systems could help FDA produce the files faster using a semi-automated approach. Extraction of drug names generally corresponds to the state-of-the-art established on other text collections, such as clinical text and the literature. Compared to previous years, the results this year indicate a significant increase in state-of-the-art for all tasks, even after accounting for differences in evaluation. The results of this evaluation have already informed the FDA and NLM collaboration on the the next steps. We hope the availability of the training and test collections will further encourage research of this imprtoant problem.

Acknowledgements

The organizers would like to thank the corpus annotators: Mark Sharp (who finalized the guidelines, annotated a large part of the test sets and provided quality assurance as NLM Special Volunteer), Phong Do, Julia Xu, and Kin Wah Fung. We would like to thank Yuqing Mao for supporting the annotation interface and ensuring its availability during annotation.

This project was primarily supported through an Inter-agency Agreement between the U.S. Food and Drug Administration (FDA) and the U.S. National Library of Medicine (NLM), part of the National Institutes of Health (NIH). This work was also partially supported by the intramural research program at the U.S. National Library of Medicine. Research reported in this publication was partially supported by the National Library of Medicine of the National Institutes of Health under award number R01LM011838. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Roberts, K., Demner-Fushman, D., Tonning, JM.: Overview of the TAC 2017 Adverse Reaction Extraction from Drug Labels Track. `https://tac.nist.gov/publications/2017/additional.papers/TAC2017.ADR_overview.proceedings.pdf`. Last accessed 30 Nov 2018
2. Goldstein, J., Jaradeh, I., Jhawar, P., Stair, T. ED Drug-Drug Interactions: Frequency & Type, Potential & Actual, Triage & Discharge. The Internet Journal of Emergency and Intensive Care Medicine **8**(2), (2004)
3. Boyce, RD., Horn, JR., Hassanzadeh, O., De Waard, A., Schneider, J., Luciano, JS., Rastegar-Mojarad, M., Liakata M. Dynamic enhancement of drug product labels to support drug safety, efficacy, and effectiveness. Journal of Biomedical Semantics **4**(5), (2013)
4. Ayvaz, S., Horn, J., Hassanzadeh, O., Zhu, Q., Stan, J., Tatonetti, NP., Vilar, S., Brochhausen, M., Samwald, M., Rastegar-Mojarad, M., Dumontier, M., Boyce, RD. Toward a complete dataset of drug-drug interaction information from publicly available sources. J Biomed Inform. **6**(55), (2015)
5. Segura-Bedmar, I., Martínez, P., Zazo, MH. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). In Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)
6. Lim, S., Lee, K., Kang, J. Drug drug interaction extraction from the literature using a recursive neural network. PLoS ONE **13**(1), (2018)
7. Fung, KW., Jao, CS., Demner-Fushman, D. Extracting drug indication information from structured product labels using natural language processing. J Am Med Inform Assoc. **20**(3), (2013)
8. Rodriguez, LM., Demner-Fushman D. Automatic Classification of Structured Product Labels for Pregnancy Risk Drug Categories, a Machine Learning Approach. AMIA Annu Symp Proc. (2015)
9. Ma, X., Hovy, E. End-to-end Sequence Labeling via Bidirectional LSTM-CNNs-CRF. in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (2016)
10. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. InProceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) 2019 Jun (pp. 4171-4186).
11. Lafferty J, McCallum A, Pereira FC. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

A Run Descriptions

The run descriptions provided by participating teams are given in Table 6.

Table 6. Run descriptions provided by participating teams

Team	Run	Description
IBMResearch	1	Pretrained Language Model based entity extraction and interaction identification. This run also includes dependency parse based post-processing.
	2	Pretrained Language Model based entity extraction and interaction identification.
	3	Language Model based entity extraction and interactions identification. Dependency parse is used for post-processing.
INK_BC	1	[H]ybrid of context and n-gram.
UTDHLTRI	1	BERT+CRF for task 1 BERT + Multi-task relation extraction for task 2/4 [S]tring matching for task 3
	2	BERT+CRF for task 1 BERT + Multi-task relation extraction for task 2/4, modified architecture [S]tring matching for task 3
	3	BERT boundary model for task 1 Multi-task BERT model for task 2 String-matching/SciSpacy for task 3 Task 4 inferred from tasks 2/3 Filtered out mentions that do not participate in a relation.
SRCB	1	Task1: NER based on BERT with additional universal transformer layers and complete automatic Data Augmentation, with average checkpoint and relative position attention. Task2: Relation identification on BERT with support sentence construction based on the meaning of each kind of interactions, with additional parser features. Task3: LTR.
	2	Task1: NER based on BERT with additional universal transformer layers and complete automatic Data Augmentation, with relative position attention but without average checkpoint. Task2: Relation identification on BERT with support sentence construction based on the meaning of each kind of interactions, with additional parser features and pre-training on NLM180 data. Task3: LTR.
	3	Task1: NER based on BERT with additional universal transformer layers and complete automatic Data Augmentation, with average checkpoint but without relative position attention. Task2: Relation identification on BERT with support sentence construction based on the meaning of each kind of interactions. Task3: Weighted features.