A Hybrid Model for Drug-Drug Interaction Extraction from Structured Product Labeling Documents

Diwakar Mahajan*
IBM Research
New York, USA
dmahaja@us.ibm.com

Ananya Poddar*
IBM Research
New York, USA
poddar@us.ibm.com

Yen-Ting Lin†
National Taiwan University
Taipei, Taiwan
b04705026@ntu.edu.tw

Abstract

Drug-Drug Interactions (DDI), which lead to adverse events, have been identified as the eighth leading cause of death in the United States (Goldstein et al., 2005). Structured Product Labeling (SPL) documents are a rich source of information for drug labels, but it is essential to convert them to discrete, coded information in order to enable automatic extraction of drug interactions. TAC 2019 DDI track defined multiple Natural Language Processing (NLP) tasks, namely concept extraction of Precipitant and SpecificInteraction (Task 1), relation extraction of interactions and their types (Task 2), concept normalization (Task 3) and generation of a global list of interactions per SPL document (Task 4). We participated in Tasks 1 and 2, where we used a combination of a novel tagging scheme, a fine-tuned transformer-based language model, and a syntactic-parse based knowledge-injected pattern matching technique. We submitted three systems for both Tasks 1 and 2. We achieved an F1-score of 65.38, 64.62 and 65.18 for Task 1, and an F1-score of 49.03, 48.33 and 48.39 for Task 2. Our system ranked first, with the **highest F1-score** for both **Tasks 1 and 2**, thus demonstrating an effective adaptation of our hybrid system on the DDI extraction tasks.

1 Introduction

Structured Product Labeling (SPL) is a Health Level Seven International (HL7) standard adopted by the U.S Food and Drug administration (FDA) in order to exchange product information. These documents express the content of human prescription drugs in an XML format. It is critical to convert the narrative text present in the SPL documents into computer-readable data, in order to enable effective deployment of drug safety information. To enable this, the FDA and the U.S. National Library of Medicine (NLM) have been working together by organizing NLP challenges such as Adverse Drug Reaction (ADR) extraction in Text Analysis Conference (TAC) 2017 (Roberts et al., 2017), followed by DDI extraction in TAC 2018 (Demner-Fushman et al., 2018) and TAC 2019. We participated in Tasks 1 and 2 of TAC 2019, described below:

- Task 1: **Concept Extraction task (NER):** Extraction of mentions, namely Precipitants and SpecificInteractions at a sentence level. **Precipitants** are defined as a drug, drug class or substance interacting with the label drug. **SpecificInteraction** (SI) is generally a disorder, or a biomedical result of the DDI. **Trigger** is defined as a trigger word or a phrase for an interaction event. However, Trigger mention type is not evaluated as a part of Task 1.

- Task 2: **Relation Extraction task (RE):** Identify interactions at a sentence level for the given Precipitant, associate it with a Trigger and classify the interactions into the following: **Pharmacokinetic** (drug-drug effects on each other), **Pharmacodynamic** (an effect of the drug combination on the organism) and **Unspecified** (general warnings of risk against combining Precipitant with label drug). Furthermore, identify the outcomes associated with the Pharmacodynamic interaction (SpecificInteraction).

Traditionally, a Named Entity Recognition (NER) system would be employed to find the mentions

*Equal Contribution
†Work done during internship at IBM Research

in Task 1 and a Relation Extraction (RE) system would be utilized to identify the type of interaction between label drug and the Precipitant in Task 2. However, for Task 2, we observed for every interaction relation, the trigger belonging to each relation type can be leveraged to indicate the relation types directly. Thus, we employed a DDI extraction system where we constructed NER models for both Tasks 1 and 2, one for each mention type (Precipitant, SpecificInteraction and interaction-specific Trigger), eliminating the need for a traditional RE method. Further, we observed that a substantial percentage of the mentions were discontinuous. Thus, we propose a novel hybrid methodology of concept representation, where we employ knowledge-injected syntactic-parse based pattern matching to reduce complexities due to a multi-model approach or label sparsity issues associated with other tagging schemes. We also evaluate language model based fine-tuning approach for these NER tasks.

The rest of this paper is organized as follows: we review previous related tasks and approaches in section 2. Next, we elaborate on the dataset, our observations and our modeling approach in section 3. In section 4, we describe our novel hybrid tagging scheme with its encoding and decoding algorithms. In section 5, we detail our architecture, followed by our submissions and results. Finally, we conclude with discussion and future work in section 6.

2 Related Tasks

Similar Drug-Drug Interaction tasks have been proposed in the past (Segura-Bedmar et al., 2011; Segura-Bedmar et al., 2013; Demner-Fushman et al., 2018).

Among these, Demner-Fushman et al. (2018) is the closest to the current TAC DDI challenge, sharing annotation guidelines as well as a part of the dataset. The best performance in TAC DDI 2018 was achieved by Tang et al. (2018), where they employed a two-step joint model consisting of CNN based encoder and RNN based decoder. They used fine-grained triggers to extract interaction relations and BIOHD tagging for NER tasks. Tran et al. (2019) proposed a multi-task learning framework designed to jointly model Tasks 1 and 2. In our previous work for TAC DDI 2018 (Dandala et al., 2018), we leveraged a different dataset of comparable definition to overcome the lack of sufficient ground truth. We were the only team that participated in all the tasks and hence, had an end-to-end system for Drug-Drug Interaction extraction.

Compared to last year, the dataset size has increased from 22 SPLs to 211 SPLs. The evaluation metrics have evolved to not include Trigger evaluation as a part of Task 1. Further, during evaluation the extracted mention text is compared rather than the span offset in the sentence. In our participation this year, we apply current state-of-the-art NER extraction techniques (Devlin et al., 2018) along with a novel hybrid tagging scheme for the NER and RE tasks.

3 Dataset & Modeling

As a part of the DDI Track, 211 SPLs were provided as training data to track participants. For Task 1, the training data contained 9048 Precipitant mentions, 2744 SpecificInteraction mentions and 5345 Trigger mentions. These mentions contained 322 (3.55%), 279 (10.17%) and 1876 (35.1%) disjoint mentions respectively. For Task 2, the dataset contained 3176 Pharmacokinetic Interactions, 4324 Pharmacodynamic Interactions and 2918 Unspecified Interactions.

In the following sections, we discuss the cleanup effort undertaken to improve the quality of the ground truth in subsection 3.1, followed by our modeling approach for Tasks 1 and 2 in subsection 3.2.

3.1 Data Cleanup

A sizable effort was spent in fixing span issues in the released ground truth annotations. Some of the inconsistencies observed are as follows:

- Mention spans expressed with an invalid begin index of -1.
- Inconsistency in annotation for a specific piece of text. For example: the first occurrence of *P-gp inhibitors* in a sentence is sometimes marked as a *Precipitant* in that sentence, but not the following occurrence in the same sentence.
- Sub-word annotation in mention, instead of the whole word following it. For example: annotation of a sub-word ***sirolimus*** in sentence

Type	Subtype	Sentence Text	Mention Text(s)
Regular	Continuous Entity	Coadministration of **antiplatelet agents** and chronic NSAID use increases the risk of bleeding.	antiplatelet agents
Irregular	Overlapping Entity	Avoid concomitant use of ELIQUIS with **P-gp** and **strong CYP3A4 inducers** as it will decrease exposure to apixaban.	P-gp | inducers, strong CYP3A4 inducers
	Disjoint Entity	As the blood pressure falls under the **potentiating effect** of LASIX, a further **reduction in dosage** or even **discontinuation** of other antihypertensive drugs may be necessary.	potentiating effect | reduction in dosage | discontinuation

The delimiter | indicates disjoint entities as expressed in ground truth.

Table 1: Entity Type Examples

*Patients receiving coadministration of ACE inhibitor and mTOR inhibitor (tem**sirolimus**, sirolimus) therapy may be at increased risk for angioedema*.

- Incorrect annotation of discontinuous mention text, in addition to its super-span in the same sentence. For example: A span such as *P-gp and strong CYP3A inhibitors* is sometimes marked as three individual Precipitants, namely; *P-gp* | *inhibitors* and *strong CYP3A inhibitors* in addition to its super-span *P-gp and strong CYP3A inhibitors*.

We developed semi-automated approaches to correct such instances based on the timely feedback received from track organizers. This helped improve the consistency of our training data, and hence the quality of our system predictions.

3.2 Modeling Approach

We propose a multi-step modeling approach to identify the mentions (Precipitant and SpecificInteraction) which form Task 1 and the interaction relation types (Pharmacodynamic, Pharmacokinetic and Unspecified) which form Task 2. In this process, we train a NER model for each mention type, namely Precipitant and SpecificInteraction. For identification of the interaction relation types, we observed that the Trigger associated with each interaction relation can be leveraged to indicate the relation type directly. Thus, we replace the relation extraction process for Task 2 with a NER task in which we model interaction-specific Triggers. The interaction-specific Triggers are of three kinds: **TRIG-K** (Pharmacokinetic Trigger), **TRIG-D** (Pharmacodynamic Trigger) and **TRIG-U** (Unspecified Trigger). We employ the tagging approach detailed in section 4 to encode and decode these entities.

Our training process results in three separate NER models, which we apply as follows:

1. Apply the **Precipitant model** to identify Precipitants in a sentence.

2. Determine if any predicted Precipitants are label drug or a variant of the label drug (e.g. generic name, drug class), and we remove such Precipitants.

3. Apply the **interaction-specific Trigger model** only for sentences which have valid Precipitants, and identify the type of interaction relation for that Precipitant.

4. In case of TRIG-D, we apply the **SpecificInteraction model** to identify the effect of the Pharmacodynamic interaction.

‡DDI Guidelines https://bionlp.nlm.nih.gov/tac2019druginteractions/DDIvalidationGuidelines.docx

‡DDI Decision Tree https://bionlp.nlm.nih.gov/tac2019druginteractions/DDI_decision_tree.xlsx

These steps mirror the human annotation process as specified in the TAC 2019 decision tree and guideline documents[‡].

4 Tagging Scheme

Typically, named entities are *regular* concepts, with a continuous sequence of words. Thus, NER systems encode annotated concepts using BIO tagging, where each token is assigned into one of the three labels: B means beginning, I means inside, and O means outside of a concept. However, BIO tagging is not sufficient for this NER task as 14% of the concepts are *irregular*. Further, we found there are two types of irregular concepts:

- **Overlapping Entity**, which is a *group* of two or more concepts that share a token or a phrase. As shown in Table 1, the two concepts *P-gp | inducers* and *strong CYP3A4 inducers* belong to a single Overlapping Entity as they share *inducers*.

- Non-overlapping **Disjoint Entity**, which is an irregular concept which has no shared tokens with any other concept. As shown in Table 1, a single concept *potentiating effect | reduction in dosage | discontinuation* forms a single Disjoint Entity.

In the following subsection 4.1, we review the popular tagging schemes currently applied on irregular concepts, and their drawbacks when applied on the more complex examples, as exemplified in Table 2. In subsection 4.2, we propose our hybrid tagging scheme, and discuss its improvements and limitations. Finally, in subsection 4.3, we detail our decoding scheme, which comprises of syntactic-parse based knowledge-injected post-processing.

4.1 Previous Work

To overcome the shortages of BIO tagging for irregular concepts, Tang et al. (2013) has suggested a BIOHD tagging scheme which works well for disjointed and overlapping concepts. In this tagging scheme there are 7 labels {B I O HB HI DB DI} defined as follows:

- HB and HI refer to tokens that are shared by multiple concepts. These tokens are the overlapped portions of disjoint concepts. These tokens or sequence of tokens are referred to as head components.

- DB and DI refer to tokens that belong to disjoint concepts, however these tokens are not shared by multiple concepts. These tokens or sequence of tokens are referred to as non-head components.

- B and I are used to label the tokens that belong to continuous concepts and,

- O refers to tokens that are outside of concepts

For decoding, it is trivial to merge continuous concepts (BIO tags). For irregular concepts, Tang et al. (2013) suggests merging head components with all other non-head components, and in absence of a head component combine all non-head components to form irregular concepts. However, the decoding process suffers from ambiguity when there are multiple occurrences of any type of irregular concept in a sentence, failing to reconstruct mentions for all categories listed in Table 2.

As a solution to this drawback, Tang et al. (2015) proposed BIOHD1234, and Li et al. (2018) proposed NerOne. In previous works, we have employed these tagging techniques for extracting Adverse Drug Reactions (Dandala et al., 2017) and extracting Drug-Drug Interactions (Dandala et al., 2018) from SPL documents, which had similar NER tasks with irregular concepts. However, we observed that these tagging schemes have an added layer of complexity, in the form of label sparsity (Tang et al., 2015) or training an additional classification submodel (Li et al., 2018).

Thus, we suggest an alternate hybrid tagging approach. During the encoding phase, we:

- identify overlapping entities from shared concepts and merge them, thereby, converting them into continuous concepts and eliminating the need for HB and HI tags

- apply DB and DI tags on non-overlapping disjoint entities

During the decoding phase, we employ syntactic-parse based knowledge-injected pattern matching to extract mentions. Our encoding and decoding process is detailed in the following sections.

Category	Sentence Text	Mention Text(s) as in the Ground Truth	Mention Text(s) after DBIO Encoding Process
Multiple† Overlapping Entities	Avoid concomitant use of ELIQUIS with **P-gp** and **strong CYP3A inhibitors**, as well as **P-glycoprotein** and **other CYP inducers**.	P-gp | inhibitors**,** strong CYP3A4 inhibitors**,** P-glycoprotein | inducers, other CYP inducers | inducers	P-gp and strong CYP3A4 inhibitors, P-glycoprotein and other CYP inducers
Multiple Overlapping & Disjoint Entities	Combined **P-gp** and **strong CYP3A inhibitors** and **other drugs that** , like XARELTO , **impair hemostasis** increases the risk of bleeding.	P-gp | inhibitors**,** strong CYP3A4 inhibitors**,** other drugs that | impair hemostasis	P-gp and strong CYP3A inhibitors**,** other drugs that like XARELTO impair hemostasis
Multiple Disjoint Entities	LASIX has a tendency to **antagonize** the skeletal muscle relaxing **effect** of tubocurarine and may **potentiate** the **action** of succinylcholine	antagonize | effect**,** potentiate | action	antagonize | effect**,** potentiate | action

The delimiter | indicates disjoint mentions as expressed in ground truth.

Table 2: Multiple Irregular Concepts in a Sentence

4.2 Hybrid DBIO Tagging Scheme

We propose a hybrid DBIO tagging scheme which has 5 labels {B I O DB DI}. These labels are applied according to the type of the entity as detailed below:

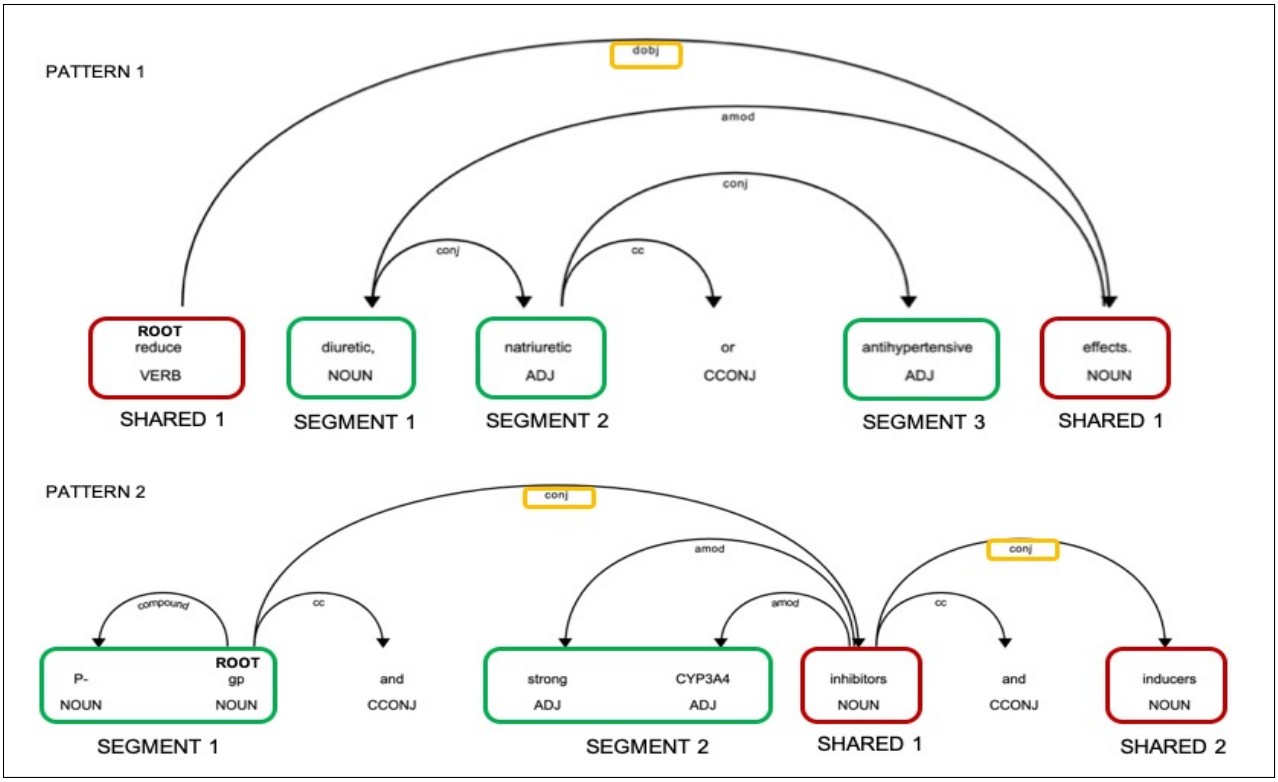
- **Continuous Entities:** We use B and I tags to label tokens belonging to continuous concepts.
- **Disjoint Entities:** For non-overlapping disjoint entities, we use DB and DI tags to label the concepts.
- **Overlapping Entities:** For an Overlapping Entity E which has a group of irregular concepts having shared token(s), we first merge the discontinuous spans of these concepts to form a merged continuous concept *m*. Start span of *m* is defined as the MIN(e_{start} in E) and end span of *m* is the MAX(e_{end} in E). We replace E with the merged and now continuous concept *m*. We use B and I tags to label *m*. Applying this technique to the overlapping entity example mentioned in Table 1, the two mentions *P-gp | inducers* and *strong CYP3A4 inducers* are merged into a single † mention *P-gp and strong CYP3A4 inducers*.
- **Others/Non Entities:** O is used to label tokens outside of the above entities.

During the decoding process, we first reconstruct all the continuous entities by combining the B and I tags, and construct a single disjoint entity by merging all disjoint components. To extract overlapping entities from continuous concepts, we apply a dependency parse-based syntactic pattern matching technique. This process is explained in section 4.3.

The proposed tagging scheme allows for a simpler post-processing step for extracting overlapping entities, without burdening the model with the added complexity of multi-training steps or label sparsity issues. Our scheme also works well for two out of three categories listed in Table 2. The four mentions in *Multiple Overlapping Entities* example are merged into a single continuous concept, while in *Multiple Overlapping and Disjoint Entities* example, three entities are merged into two. These merged concepts are successfully reconstructed into the required mentions after our decoding process.

While our tagging scheme works well for these

†The number of overlapping entities is defined by the number of shared chunks.



Pattern 1 and 2 serve as the most effective patterns on SpecificInteraction and Precipitant, respectively

Figure 1: Dependency parse-based syntactic patterns

scenarios, our system is not completely free from ambiguity. In case of category *Multiple Disjoint Entities* present in the same sentence, the decoding process does not work. However, this scenario accounted for less than 4% of the observed instances in our training data, resulting in a relatively low information loss which is a favourable trade-off for a more lightweight effective tagging technique.

4.3 Syntactic Pattern Matching based Decoding System

We employ a knowledge-injected syntactic pattern matching methodology to extract overlapping irregular entities from merged concepts. First, given a continuous span, we analyze the text to identify if it is an irregular concepts, based on:

- the presence of a conjunction *CCONJ* in its dependency parse tree
- the presence of a drug-mention (drug, namely the label drug, its variant or Precipitant) in its tokens

We use SpaCy parser (Honnibal and Johnson, 2015) to generate a dependency parse tree for each mention text, and inject each token corresponding to a drug-mention with *isDrug* = *True*. A drug-mention knowledge base is constructed by extracting the generic name and drug class for each label drug and Precipitant using RxNorm (Liu et al., 2005) as a resource.

Next, we extract constituents from the identified irregular concepts. We observe that overlapping entities are expressed in a limited number of ways, and thus we were able cover 89% of these entities with a few syntactic patterns. This technique involves two steps:

- Step 1: Identification of **shared** chunks, and **segment** chunks from the continuous span, where the tokens in the shared chunks are shared across the irregular concepts, while the tokens of the segment chunks are unique to each.
- Step 2: **Merging** each shared chunk with each

segment chunk to form discontinuous spans.

For example, for the continuous mention span *P-gp and strong CYP3A inhibitors*, the token *inhibitors* forms a shared chunk, *P-gp* is the first segment, and *strong CYP3A* is the second segment. Once identified, the shared chunk is now merged with each segment chunk to form two discontinuous spans namely, *P-gp* | *inhibitors* and *strong CYP3A inhibitors*.

Figure 1 shows two of our most commonly applied patterns. Specifically, **Pattern 1** is applied on SpecificInteraction, where the *root* of the sentence, and its *dobj* (direct object) are extracted to form a shared chunk, and each *conj* (co-ordinated token) form individual segment chunks. **Pattern 2** is applied on Precipitant. Here, each *conj* forms individual shared chunks, while the *root* and *lefts* (left immediate children) of the shared chunks form the individual segment chunks. It should be noted that in each chunk, the *compound* for each token is also consumed.

Further, we follow the guidelines to remove any drug-mention tokens that are a part of the predicted SpecificInteraction. For example: In a predicted SI text *reduce* ***lithium****'s renal clearance*, the token *lithium* where *isDrug* = *True*, is simply removed, resulting in the extraction of a disjoint entity *reduce* | *renal clearance*.

5 Experiments & Results

Recently, transformer-based language models have achieved state-of-the-art results in several NLP tasks including Named Entity Recognition (Devlin et al., 2018). In this process, the out-of-the-box pretrained language model BERT (Devlin et al., 2018), is fine-tuned on the target task (NER task in our case) and thus applies the learned encoded information from pretraining on a huge corpus. Further, BERT breaks down input words into sub-word tokens referred to as WordPieces (Schuster and Nakajima, 2012). These WordPieces are generated via statistical analysis on a large corpus rather than using a morphological lexicon. Since we wish to employ the pre-trained language model, we continue to use the WordPiece vocabulary utilized by BERT for the fine-tuning task.

We split the provided dataset (211 SPLs) into 75% (train), 15% (validation) and 10% (blind) of the data for submissions 1 and 2; and 80% (train), 20% (validation) for submission 3. We perform a 5-fold cross-validation for each mention type, thereby training 15 models (5 models per mention type).

Parameter	Value
Learning Rate	1e-5
Number of epochs	20
Batch size	16
Dropout	0.1
Optimizer	Amsgrad

Table 3: Experimental setup

Submission	Train Size	Task 1	Task 2
1	90%	**0.6538**	**0.4903**
2	90%	0.6462	0.4833
3	100%	0.6518	0.4839

Train size refers to the proportion of training data used per submission.

Table 4: Official Test Data F1-Score

5.1 Sequence Labeling Model

We fine-tune the **24-layer, 1024-hidden, 16-head BERT-Large, Cased model (Whole Word Masking)** (Devlin et al., 2018) for each mention extraction task i.e. for Precipitant, SpecificInteraction and Trigger. We use BERT to extract the contextualized embedding for each token, and add a fully connected layer on top of BERT to classify each token into the mention type. Our experimental setup is shared in Table 3.

The weights in all the layers of the model are updated during training. The tokenization method follows Devlin et al. (2018), but for GPU memory concerns, we set the token limit (after WordPiece tokenization) to 180 during training and 512 during inference. We assign an O tag to each token outside our limit. For a token split into several sub-tokens after tokenization, we only consider the first sub-token and ignore the rest, i.e. we only calculate the loss for the first sub-token during training and take the prediction of the first sub-token as the prediction for the whole token.

Finally, we employ a **Max-Voting system** for

Category	Precision	Recall	F1-Score
Task 1			
Typed-Mentions*	**0.734**	**0.589**	**0.6538**
Precipitant	0.748	0.665	0.704
SpecificInteraction	0.664	0.358	0.466
Task 2			
Relation Type	**0.809**	**0.643**	**0.717**
Pharmacokinetic	0.866	0.65	0.742
Pharmacodynamic	0.893	0.64	0.71
Unspecified	0.784	0.65	0.711
Typed-Interactions*	**0.583**	**0.423**	**0.4903**
Pharmacokinetic	0.711	0.568	0.632
Pharmacodynamic	0.563	0.365	0.434
Unspecified	0.551	0.429	0.483

* indicates the primary evaluation metrics.
Rows in bold specify the aggregated scores for the following rows.

Table 5: Breakdown of Test Data Scores for Best Submission

each mention type leveraging models generated during the 5-fold cross-validation. This helps minimize the variance in the predictions, balancing the observed inconsistency in the training data.

5.2 Results & Analysis

Table 4 shows our official results, and Table 5 shows a further breakdown of these results on the official test data.

Our Submissions 1 and 3 employ the complete syntactic dependency parse-based system, as explained in Section 4.3, as a post-processing reconstruction step. Our Submission 2 applies only the knowledge-injected token removal step. Our Submission 1 achieved the highest F1-score for Tasks 1 and 2.

In Table 5, Typed-Mentions is evaluated based on the mention text, and Typed-Interactions is evaluated based on the interaction type, its associated Precipitant text and effect, if present i.e. SpecificInteraction text for Pharmacodynamic Interaction. Relation Type is evaluated based on the frequency of an interaction type for a given sentence.

Overall, our system is more precision-oriented. Based on a deeper error analysis, our common error categories and observations are as follows:

- **Issues in exact Precipitant text extraction**: 29% of the total number of predicted Precipitants have an overlap with the gold standard. This includes cases such as,
gold = *insulin lispro* and prediction = *insulin lispro product*; gold = *serotonergic | drugs* and prediction = *drugs that affect the serotonergic neurotransmitter system*; gold = *preparations containing sulfur* and prediction = *sulfur*. Additionally, the label drug variant removal step introduced a few false negatives. These issues further cascade, thereby negatively impacting Task 2 primary evaluation metrics.

- **Complexities in SpecificInteraction definition**: Given that SI text is generally long, we found a partial match between the gold standard and predicted text in 17% of the total instances. For example: gold = *electrocardiographic changes | hypokalemia* and prediction = *hypokalemia*. As seen in Table 5, our SI model suffers from a recall issue. On further analysis, we observe that 11% of total SI instances are discarded due to the absence of Precipitant for that sentence. Next, we observe that TRIG-D recall is higher, thereby indicating that a joint modeling between TRIG-D and SI could benefit the overall recall for SI. This, in combination with the complexity of SI definition, made the overall SI extraction task more difficult. It should be noted that, while the num-

ber of training instances for SI is only one-third as that for Precipitant, the number of irregular concepts is three times higher.

- **Errors in Precipitant cascading to Typed-Interaction Evaluation**: The errors in Precipitant reported earlier further cascade to errors in Typed-Interactions as shown in Table 5. This is reflected by the difference in scores between Relation Type, where we perform considerably well, and Typed-Interaction for each interaction type. We perform the worst in Typed-Pharmacodynamic Interaction, since the errors associated with SpecificInteractions are further cascaded into the same.

6 Conclusion & Future Work

In this work, we describe our participation TAC Drug-Drug Interaction Challenge 2019, where we participate in Tasks 1 and 2. We demonstrate the application of state-of-the-art transformer based techniques for extracting mentions and relations. We also propose a novel hybrid tagging methodology for irregular concepts which is lightweight and overcomes several limitations of other tagging schemes. Our system proves to be effective as we achieve the highest F1-score in the challenge. Our future directions include:

- Employing external knowledge bases (e.g. UMLS, RxNorm) for identifying SpecificInteraction and Precipitant mentions during the NER prediction in addition to the post-processing steps.

- Injecting knowledge of Precipitant and Interaction-specific Trigger while training SpecificInteraction NER model.

- Analyzing techniques to handle multiple disjoint entities to have a more robust tagging scheme.

- Analyzing the effect of using a clinical data specific WordPiece vocabulary, instead of general domain WordPieces while pretraining and fine-tuning transformer based language models.

Acknowledgments

We would like to thank Dr. Ching-Huei Tsou (IBM Research) and Prof. Weichung Wang (National Taiwan University, Taiwan) for supporting this work. We would also like to thank Dr. Jennifer Liang (IBM Research) for her insightful feedback, Yu-Cheng Wang (National Tsing Hua University) for his detailed data-analysis.

References

Bharath Dandala, Diwakar Mahajan, and Murthy V Devarakonda. 2017. Ibm research system at tac 2017: Adverse drug reactions extraction from drug labels. In *TAC*.

Bharath Dandala, Diwakar Mahajan, and Ananya Poddar. 2018. IBM Research system at TAC 2018: Deep learning architectures for drug-drug interaction extraction from structured product labels. In *Proceedings of the 2018 Text Analysis Conference (TAC 2018)*.

Dina Demner-Fushman, Kin Wah Fung, Phong Do, Richard D Boyce, and Travis R Goodwin. 2018. Overview of the tac 2018 drug-drug interaction extraction from drug labels track.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

JN Goldstein, IE Jaradeh, P Jhawar, and TO Stair. 2005. ED drug-drug interactions: Frequency & type, potential & actual, triage & discharge. *The Internet Journal of Emergency and Intensive Care Medicine*, 8(2).

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Fei Li, Meishan Zhang, Bo Tian, Bo Chen, Guohong Fu, and Donghong Ji. 2018. Recognizing irregular entities in biomedical text via deep neural networks. *Pattern Recognition Letters*, 105:105–113.

Simon Liu, Wei Ma, Robin Moore, Vikraman Ganesan, and Stuart Nelson. 2005. RxNorm: Prescription for electronic drug information exchange. *IT Professional*, 7(5):17–23.

Kirk Roberts, Dina Demner-Fushman, and Joseph M Tonning. 2017. Overview of the tac 2017 adverse reaction extraction from drug labels track. In *TAC*.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152. IEEE.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA. Association for Computational Linguistics.

Isabel Segura-Bedmar, Paloma Martinez, and Daniel Sanchez-Cisneros. 2011. The 1st DDIExtraction-2011 challenge task: Extraction of Drug-Drug Interactions from biomedical texts. volume 2011, pages 1–9.

Buzhou Tang, Qingcai Chen, Xiaolong Wang, Yonghui Wu, Yaoyun Zhang, Jingqi Wang, and Hua Xu. 2015. Recognizing disjoint clinical concepts in clinical text using machine learning-based methods. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2015:1184–93.

Buzhou Tang, Yonghui Wu, Min Jiang, Joshua C. Denny, and Hua Xu. 2013. Recognizing and encoding discorder concepts in clinical text using machine learning and vector space model. In *CLEF*.

Siliang Tang, Qi Zhang, Tianpeng Zheng, Mengdi Zhou, Zhan Chen, Lixing Shen, Xiang Ren, Yueting Zhuang, Shiliang Pu, and Fei Wu Wu. 2018. Two step joint model for drug drug interaction extraction. In *Proceedings of the 2018 Text Analysis Conference (TAC 2018)*.

Tung Tran, Ramakanth Kavuluru, and Halil Kilicoglu. 2019. A multi-task learning framework for extracting drugs and their interactions from drug labels. *arXiv preprint arXiv:1905.07464*.