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

Diwakar Mahajan[1]⋆, Ananya Poddar[1]⋆, Yen-Ting Lin[2]

[1]IBM Research New York
[2]National Taiwan University Taipei

⋆Equal contributors, Presenters

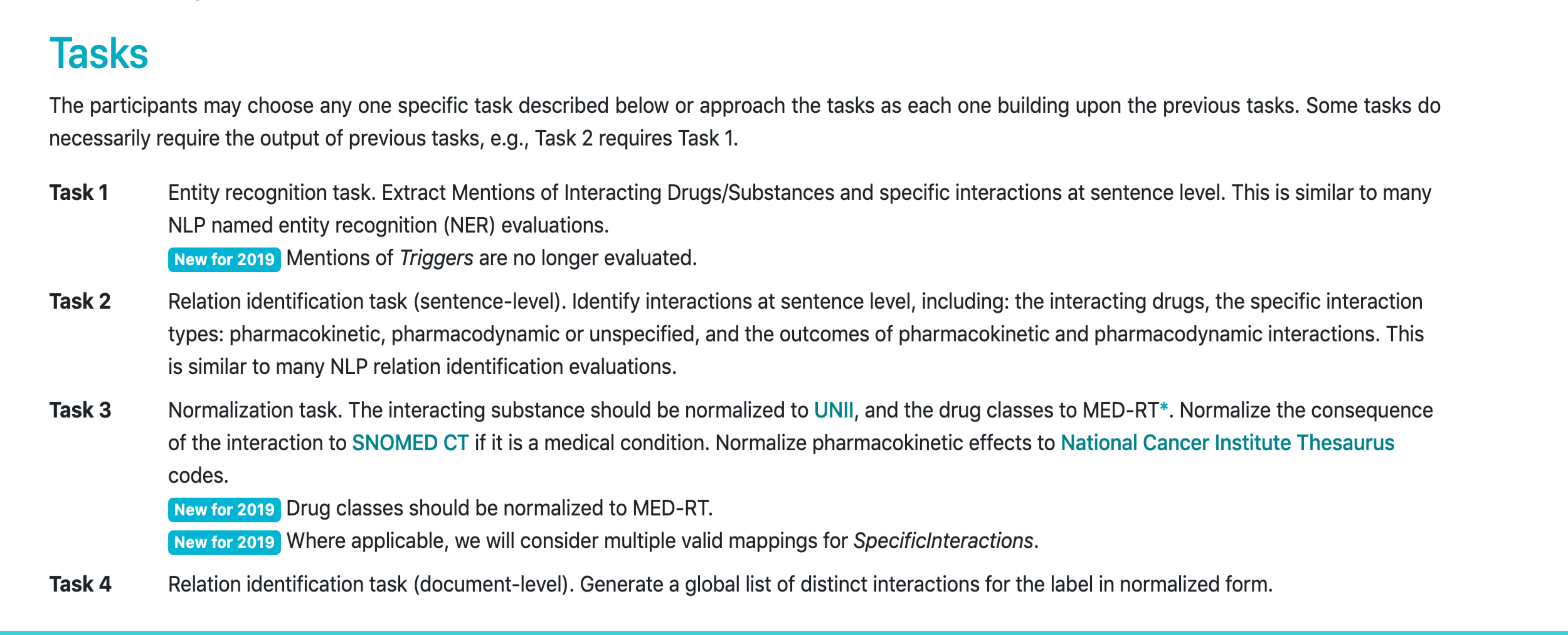
Introduction

DDI, a common cause of ADR, is the eight leading cause of death in the United States

Need for Transforming SPL into structured information

TAC 2019 DDI Challenge involves identification of drugs and interactions between them from SPLs.

End goal is to have an aggregated list of interactions.



Tasks

The participants may choose any one specific task described below or approach the tasks as each one building upon the previous tasks. Some tasks do necessarily require the output of previous tasks, e.g., Task 2 requires Task 1.

Task 1 Entity recognition task. Extract Mentions of Interacting Drugs/Substances and specific interactions at sentence level. This is similar to many NLP named entity recognition (NER) evaluations.

New for 2019 Mentions of *Triggers* are no longer evaluated.

Task 2 Relation identification task (sentence-level). 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*. Normalize the consequence of the interaction to SNOMED CT if it is a medical condition. Normalize pharmacokinetic effects to National Cancer Institute Thesaurus codes.

New for 2019 Drug classes should be normalized to MED-RT.

New for 2019 Where applicable, we will consider multiple valid mappings for *SpecificInteractions*.

Task 4 Relation identification task (document-level). Generate a global list of distinct interactions for the label in normalized form.

Task Description

Annotation Type	*Number of Annotations*	*Number of Irregular Annotations*	*Example*	*Description*
Precipitant	9048	322 (3.5%)	Changes in blood pressure must be carefully monitored when LASIX is used with other ***antihypertensive drugs.***	Interacting substance with a label drug i.e. drug, drug class or non-drug substance.
Specific Interaction	2744	279 (10.2%)	***Changes in blood pressure*** must be carefully monitored when LASIX is used with other antihypertensive drugs.	Result of interaction.
Trigger	5345	1876 (35.1%)	Changes in blood pressure must be carefully ***monitored*** when LASIX is used with other antihypertensive drugs.	Trigger word or phrase for an interaction event.

Annotation Type	*Number of Annotations*	*Example*	*Description*	*Effect*
Pharmacokinetic	3176	Lithium generally should not be given with diuretics because they *reduce* lithium's *renal clearance.*	Indicated by triggers, involves effect on absorption, distribution, metabolism and excretion of interacting drug.	Contains effect attribute; Effect is always a NCI code (C#####)
Pharmacodynamic	4324	***Changes in blood pressure*** must be carefully *monitored* when LASIX is used with other antihypertensive drugs.	Indicated by triggers and specific Interactions, is the effect of the drug combination on the organism.	Contains effect attribute; Effect is always of mention type specific interaction
Unspecified	2918	*Avoid* use of aliskiren with VASOTEC in patients with renal impairment.	Indicated by triggers, are general warning of risk against combining label drug with precipitant.	Does not contain effect attribute

Task Dataset

211 Structured Product Labeling (SPL) documents were created in the gold-standard format by NLM and FDA.

Triggers were not evaluated in Task 1.

Challenges

- A large proportion of ***irregular entities*** in data.

- Need for ***complex tagging schemes*** to handle irregular entities.

- Ground Truth Quality Issues
  - Mention spans expressed with an invalid begin index of -1.

    ```
    AGGRASTAT 2432   -1 16  “thrombocytopenia”
    ```

  - Inconsistency in annotation for a specific piece of text.

    ```
    Revatio   2673   “Concomitant PDE-5 inhibitors: Avoid use with Viagra or other PDE-5 inhibitors.”
    ```

  - Annotation of a sub-word, instead of a whole-word in sentence.

    ```
    Lotensin 3852   “Patients receiving coadministration of ACE inhibitor and mTOR inhibitor (temsirolimus, sirolimus) therapy may be at increased risk for angioedema”.
    ```

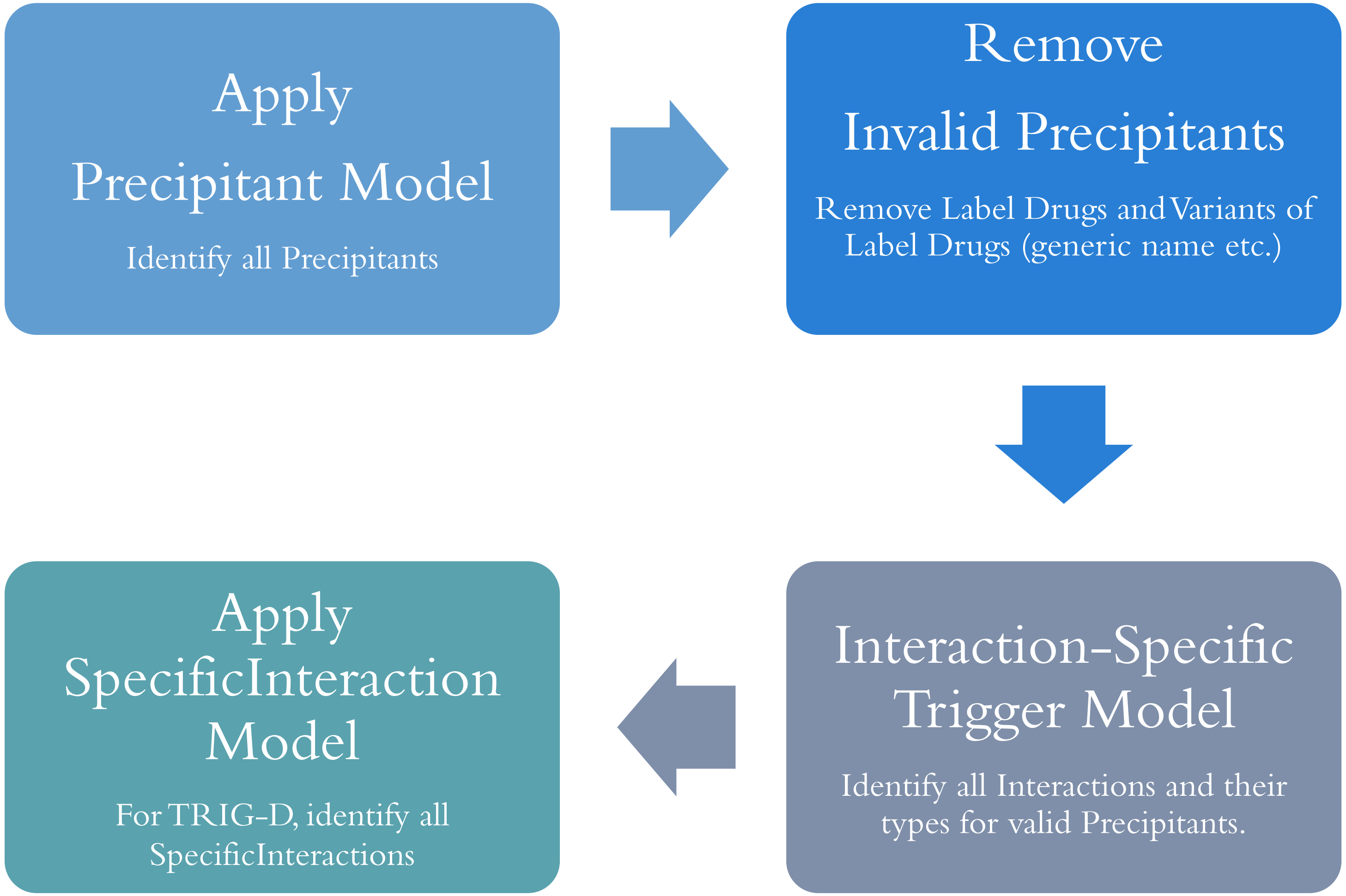
Methodology - I

Ground Truth (GT) corrections

- Timely feedback from Organizers helped improve quality of GT and thus our models.
- Developed semi-automated ways of correcting Ground Truth.

Replacing Relation Extraction (Task 2) with a NER Task - Interaction-Specific Trigger

- ***Trigger*** belonging to each relation type can be leveraged to indicate ***Interaction Types*** directly.
- The ***Interaction-specific Triggers*** are of three kinds: ***TRIG-K*** (Pharmacokinetic Trigger,
- ***TRIG-D*** (Pharmacodynamic Trigger) and
- ***TRIG-U*** (Unspecified Trigger).
- Thus we have ***3 NER models – Precipitant, SpecificInteraction, Interaction-specific Trigger***



Apply
Precipitant Model

Identify all Precipitants

Remove
Invalid Precipitants

Remove Label Drugs and Variants of Label Drugs (generic name etc.)

Interaction-Specific
Trigger Model

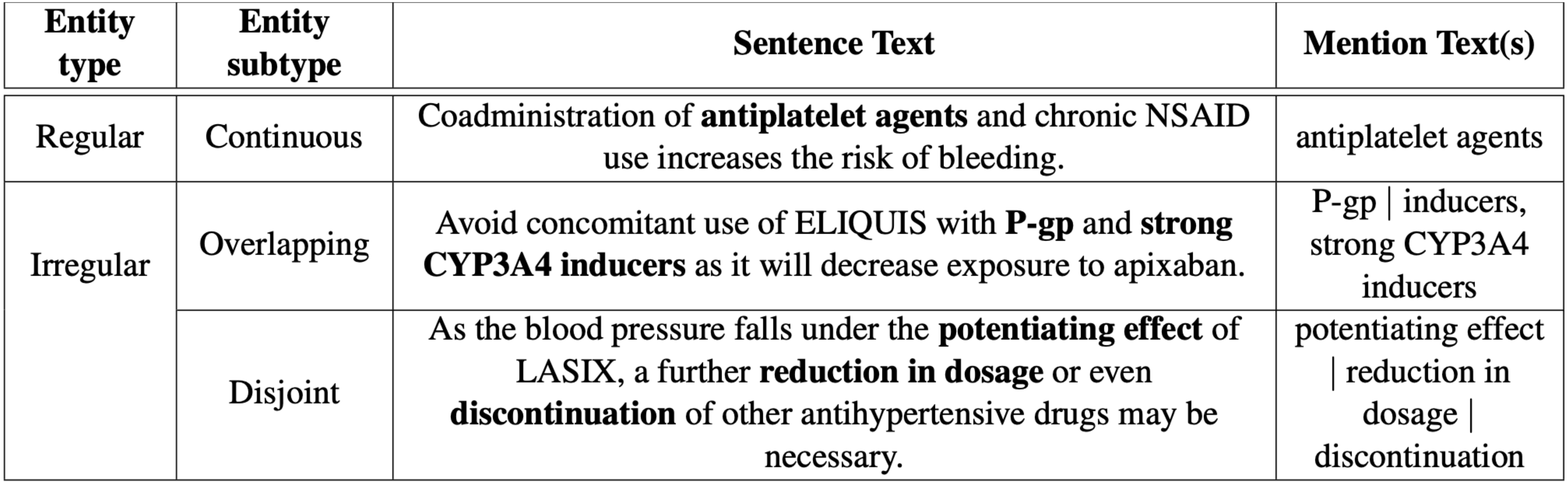
Identify all Interactions and their types for valid Precipitants.

Apply
SpecificInteraction
Model

For TRIG-D, identify all SpecificInteractions

Methodology – II

Tagging Scheme
Types of Entities



Entity type	**Entity subtype**	**Sentence Text**	**Mention Text(s)**
Regular	Continuous	Coadministration of **antiplatelet agents** and chronic NSAID use increases the risk of bleeding.	antiplatelet agents
Irregular	Overlapping	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	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.

Entity Tagging Previous Work

- ***BIOHD*** scheme has 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

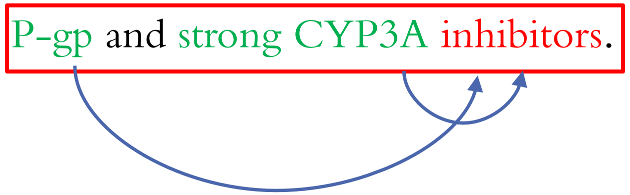
- For multiple irregular entities, ***HDBIO1234*** and ***NerOne*** have been proposed, but they suffer from label sparsity and complex multi-model approach.

Proposed Hybrid Tagging Scheme - Encoding

- ***Continuous Entities***: Use ***B*** and ***I*** tags to label tokens belonging to continuous concepts.

- ***Overlapping Entities***: For a set of entities `e` which have shared token(s), merge the discontinuous spans of these entities to form a merged continuous entity `m`. Replace the set `e` with the merged and now continuous concept `m`*.* Use ***B*** and ***I*** tags to label tokens belonging to continuous concepts.

- E.g. P-gp and strong CYP3A inhibitors.



- ***Disjoint Entities***: For non-overlapping disjoint entities, we use DB and DI tags to label the concepts.

- ***Others/Non Entities***: O is used to label tokens outside of the above entities.

- ***Continuous Entities:***
  Coadministration/O of/O **antiplatelet/B_PREC agents/I_PREC** and/O chronic/O NSAID/O use/O increases/O the/O risk/O of/O bleeding/O.

- ***Overlapping Entities:***
  Avoid/O concomitant/O use/O of/O ELIQUIS/O with/O **P-gp/B_PREC** and/**I_PREC strong/I_PREC CYP3A/I_PREC inhibitors/I_PREC**.

- ***Disjoint Entities:***
  As/O the/O blood/O pressure/O falls/O under/O the/O **potentiating/DB_SI effect/DI_SI** of/O LASIX/O, a/O further/O **reduction/DB_SI in/DI_SI dosage/DI_SI** or/O even/O **discontinuation/DB_SI** of/O other/O antihypertensive/O drugs/O may/O be/O necessary/O.

Examples - Encoding

Proposed Hybrid Tagging Scheme – Decoding & Post-Processing

- Concept Decoding:
  - Continuous concepts: Using BIO Tags
  - Disjoint concepts: Merge non-overlapping {DB DI} components

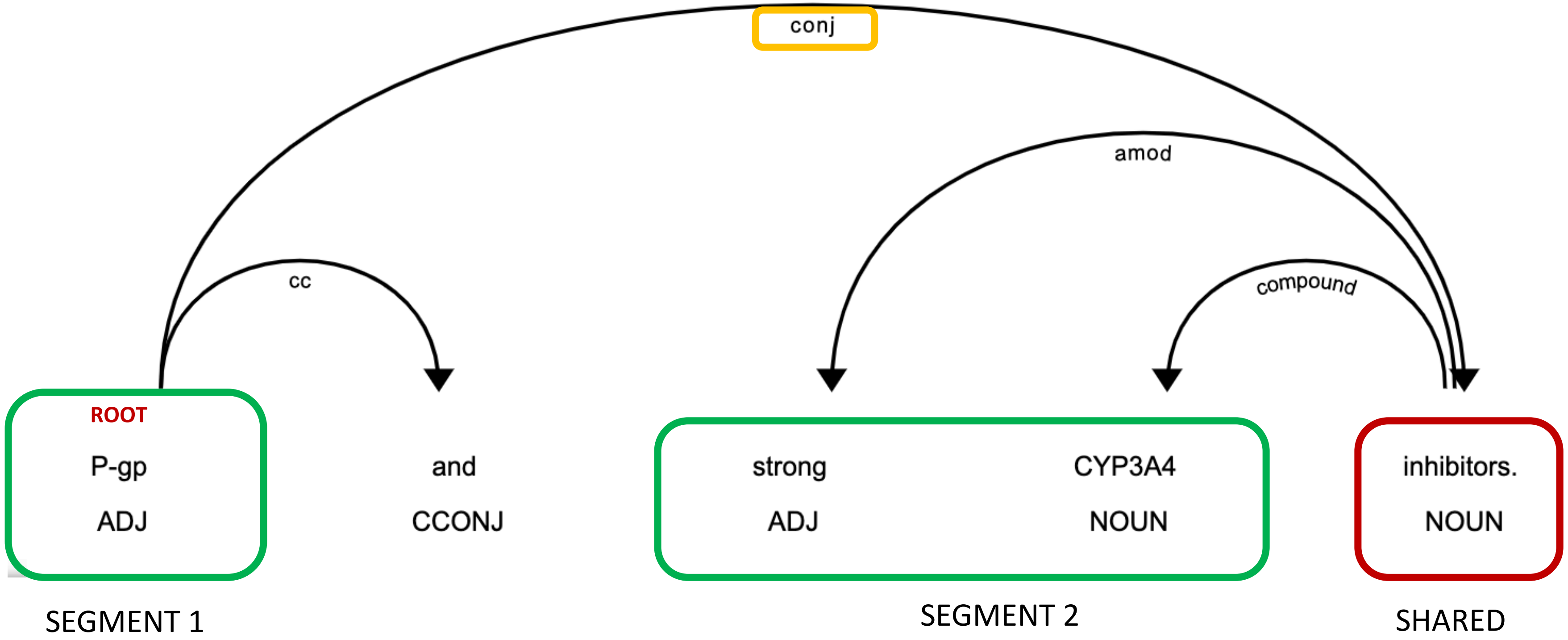
- We use SpaCy model *en_core_web_lg* for:
  - Sentence segmentation
  - Tokenization
  - Syntactic features

- We leverage Dependency Parse Trees to:
  - Identify ***frequent patterns*** to ***extract overlapping entities*** from merged entities
  - Few syntactic patterns were constructed to cover 89% of total overlapping entities

Decoding – Extracting Overlapping Entities

- Step 1: Identify ***SHARED CHUNKS*** and ***SEGMENT CHUNKS*** from span.
  - ***SHARED CHUNKS***: Tokens are shared across irregular mentions
  - ***SEGMENT CHUNKS***: Tokens are unique to each irregular mention
- Step 2: Merge each shared chunk with each segment chunk

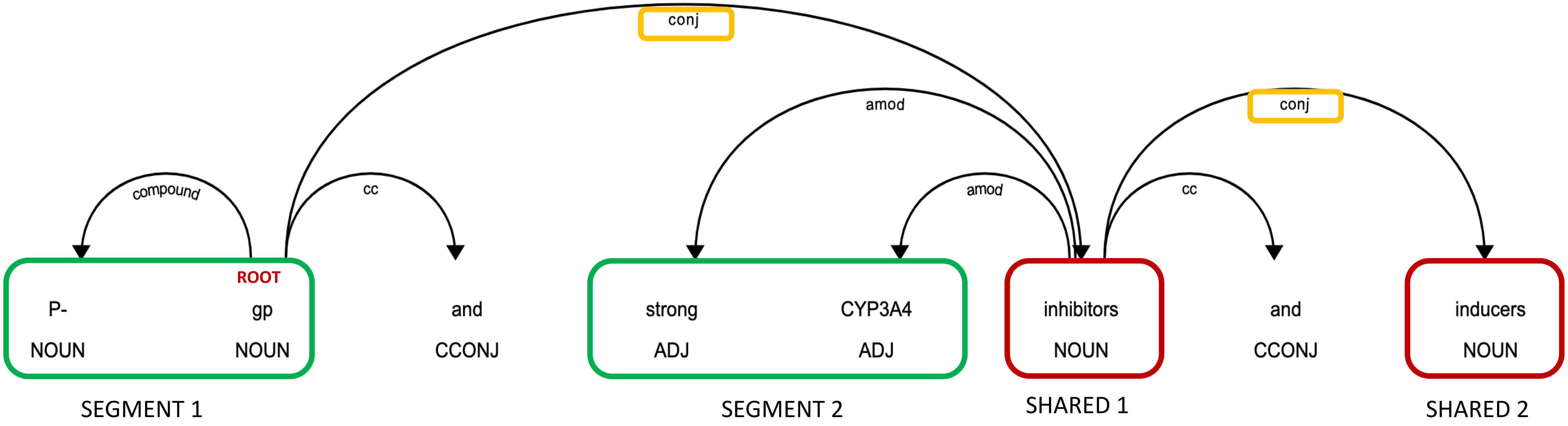
PATTERN 1 – Example 1 *(Precipitant)*



P-gp | inhibitors,

strong CYP3A4 inhibitors

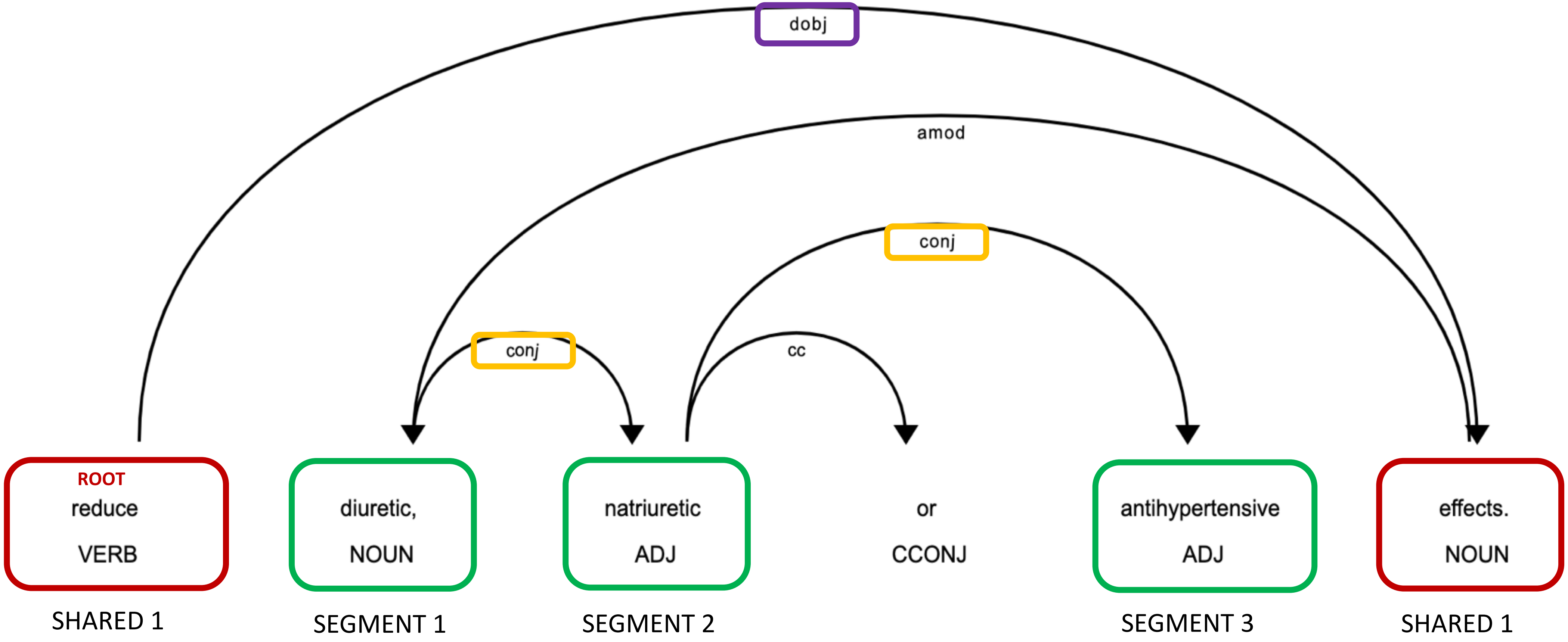
PATTERN 1 – Example 2 *(Precipitant)*



P-gp | inhibitors, P-gp | inducers,

strong CYP3A4 inhibitors, strong CYP3A4 | inducers

PATTERN 2 – Example 1 *(SpecificInteraction)*



reduce diuretic | effects, reduce | natriuretic | effects,

reduce | antihypertensive effects

Decoding- Drug clean-up

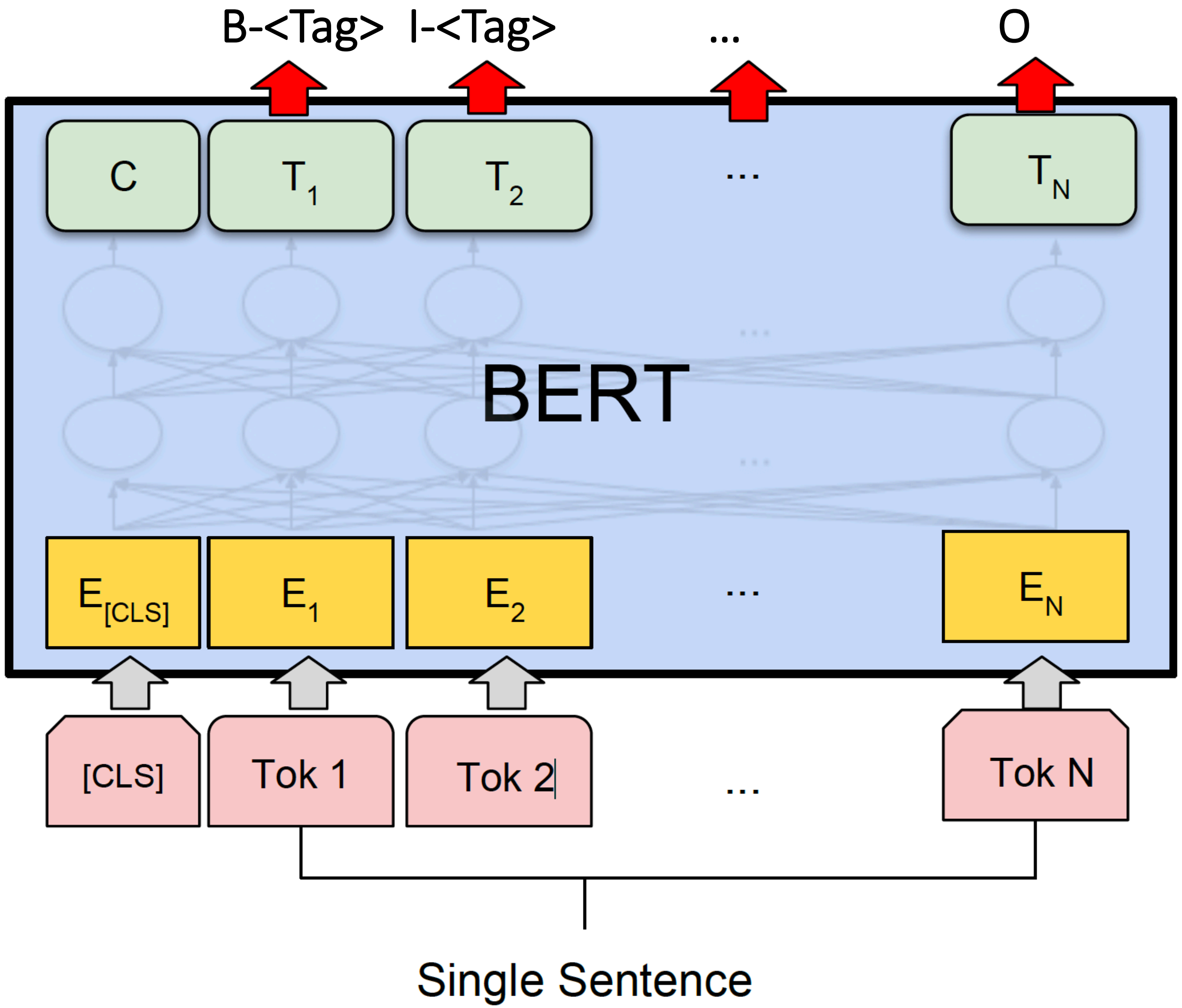
- We use ***RxNorm*** to construct a medication knowledge base
- This contains *drug class* and *generic name* of label drug

reduce lithium's renal clearance

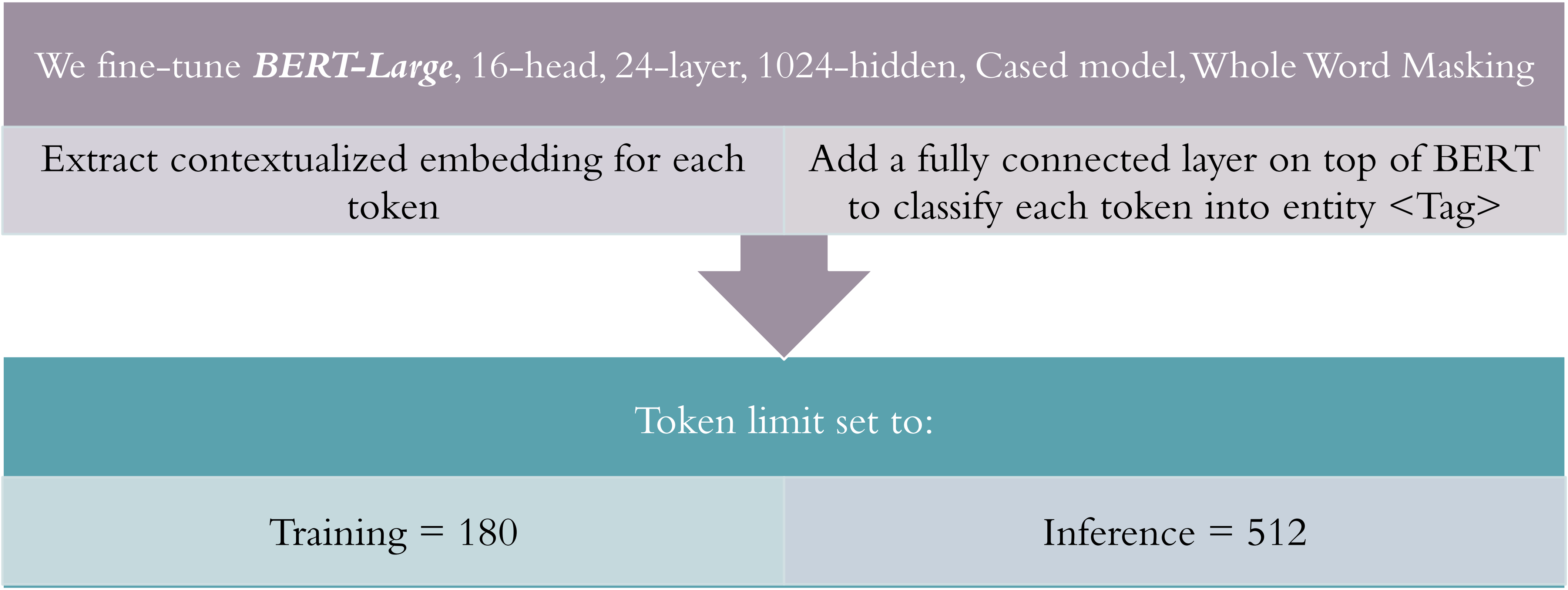
reduce | renal clearance

Sequence Labeling Method

DIAGRAM FROM DEVLIN ET. AL. (2018)



Sequence Labeling Method



We fine-tune ***BERT-Large***, 16-head, 24-layer, 1024-hidden, Cased model, Whole Word Masking

Extract contextualized embedding for each token

Add a fully connected layer on top of BERT to classify each token into entity <Tag>

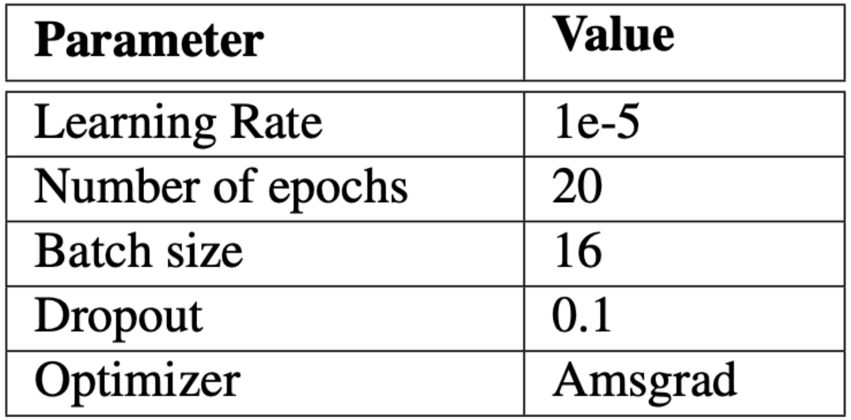
Token limit set to:

Training = 180

Inference = 512

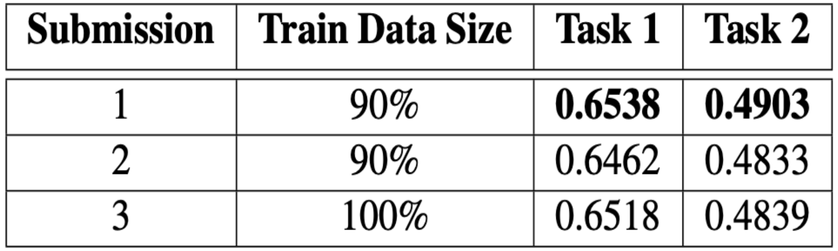
Experimental Setup

- We split the provided dataset (211 SPLs) into train, validation and blind as 75%, 15% and 10% of the data respectively.
- We perform a 5-fold cross-validation for each entity type, thereby training 15 models (5 models per entity type).
- We employ a ***Max-Voting*** system.

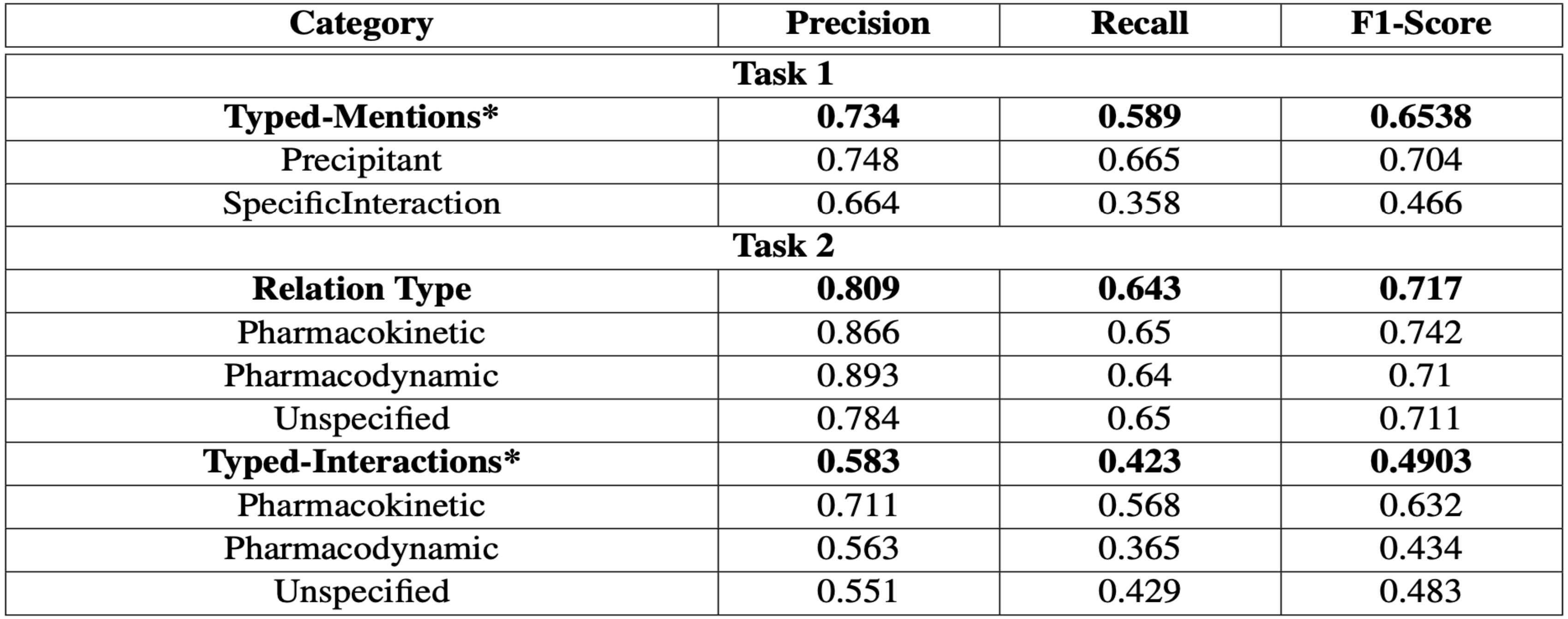


Parameter	**Value**
Learning Rate	1e-5
Number of epochs	20
Batch size	16
Dropout	0.1
Optimizer	Amsgrad

Experimental Setup



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



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

Results

Issues in Precipitant Extraction

- 29% False Negative and False Positives for Precipitants overlap
  - insulin lispro & insulin lispro product
  - serotonergic | drugs & drugs that affect the serotonergic neurotransmitter system
- Label drug variant removal step causes False Negatives

Complexities in SpecificInteraction Extraction

- 17% False Negative and False Positives for SI overlap
  - electrocardiographic changes | hypokalemia & hypokalemia
- 11% of SI discarded due to missing Precipitant
- Nuances of SI definition

Mention Extraction Errors cascading to Typed-Interaction

- Errors in Precipitant & SI cascade to errors in Typed-Interactions
- Significant difference between Relation Type and Typed-Interaction scores
- Low performance in Typed-Pharmacodynamic Interaction because of missing Precipitant & SI

Discussion

Future Work

Use external knowledge base (e.g. UMLS, RxNorm)

Joint modeling of Precipitant, SI and Interaction-Specific Trigger

Handle multiple disjoint entities

Handle nuances of data representation (e.g. better sentence segmentation, tokenization)

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 and detailed feedback.

References

[1] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991, 2015.

[2] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In Proceedings of EMNLP.

[3] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. Attention-based bidirectional long short-term memory networks for relation classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), volume 2, pages 207–212, 2016.

[4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[5] Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. arXiv preprint arXiv:1603.01354 .

[6] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv:1412.6980

[7] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recog- nition. arXiv preprint arXiv:1603.01360

QUESTIONS?