Relation Extraction: ADE-Drug and Reason-Drug using SciBERT with Positional Tagging

Prabal Bijoy Dutta

Task Overview: Obtain the established relation between a Drug with its Reason/ADE

- Approach: SciBERT tokenizer and model with positional tagging is in the input is taken as the base model, which is then concatenated it with a fully connected neural network with dropout of 30%, and trained on n2c2 dataset to develop a relation-extraction model
- What's New: Positional Tagging is used in SciBERT, which helps learn the values of the newly added tokens to extract relations.
- Attained an accuracy (F1-score) of **58%** for Reason-Drug and **78%** for ADE-Drug

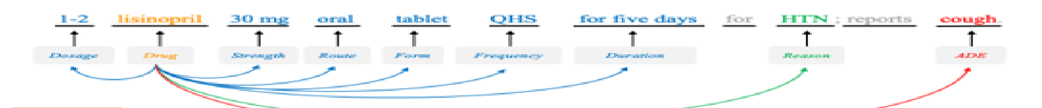
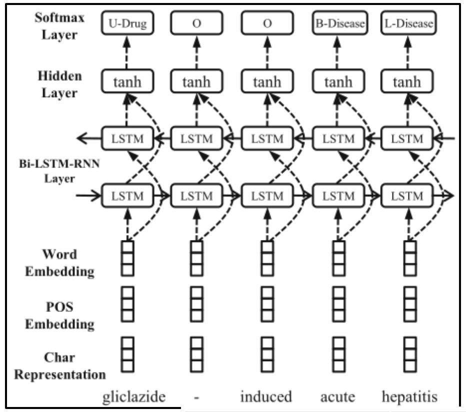


Fig.1

Text	Annotations	Examples	
He received Solu-Medrol and Levofloxacin. In the Emergency Department, a femoral line was placed, given his low blood pressure and he was started on Dopamine for the hypotension.	T9 Drug Dopamine T10 Reason hypotension R5 Reason-Drug T10:T9	Positive Reason: Drug Hypertension - Dopamine	Negative Reason: Drug Levofloxacin-hypotension.

Baseline Method:

- Introduced a joint entity model that simultaneously processes the task of entity recognition and relation classification.
- Used CNN-based character-level encoding with word embedding and POS embedding as input.
- trained two concatenated Bi-LSTM-RNN, first for identifying entities and the other for classification



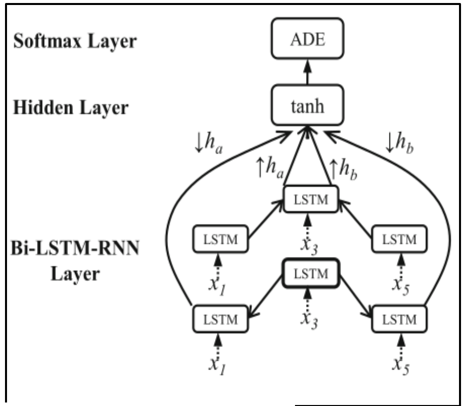


Fig.2

ADE DRUG EXTRACTION			
JOINT-ENTITY	Precision	Recall	F1-Score
Training Set	64.4	63.3	63.9
Testing Set	64	62.9	63.4

Proposed Method:

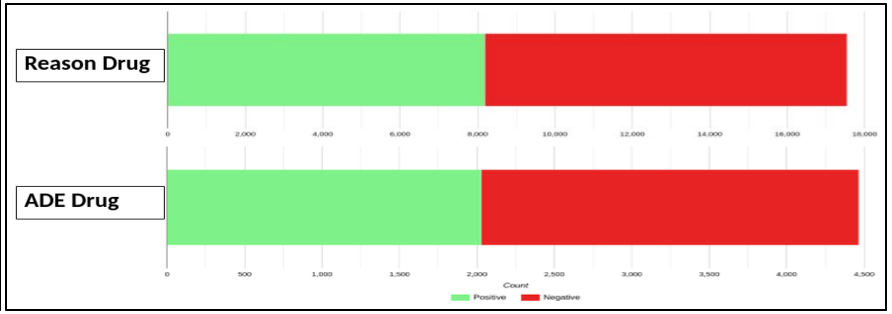
Preprocessing I
- Extracted un-structured corpus from the n2c2 dataset and assigned label as per presence of ADE/Reason-Drug

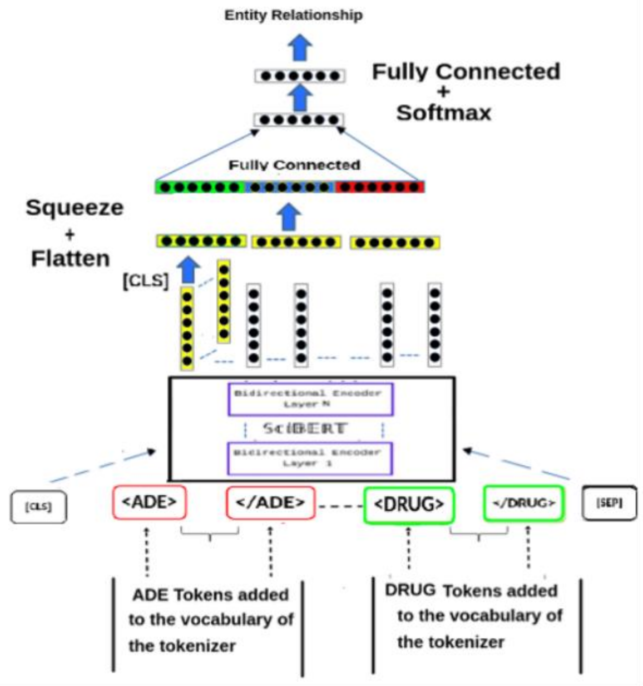
Preprocessing II
- Generated Negative Samples to attain some kind of balanced-dataset
- Cleaned the corpus removing '*Λ' etc.

Positional Tagging
- Incorporated new tokens in the SciBERT tokenizer
- Positionally tagged the relation pairs in the corpus.

Model Formation
- Pre-trained SciBERT model is appended with feed-forward layers to through the CLS node to classify relation

Hyper-parameter
- Token length: 160
- Optimizer:AdamW(2e-5)
- Loss: CrossEntropyLoss
- Drop Out: 30%





Method Details:
- Excess token is padded with zeroes.
- attention mask for padded tokens are made zero while corpus and relation are ones.

Observations:

- Observation 1: Positional Tagging has improved the Precision in both the cases, the improvement is seen more in the ADE-Drug Model. (Table 1)
- Observation 2: Recall has improved for the positive labels, in both the models (ADE/Drug & Reason/Drug). (Table 1)
- Observation 3: Improvement in Precision score has overall improved the total F1-Score in both the model. (Table 1)
- Observation 4: F1 Score for either of the positive and negative labels have improved, which is mainly due to change in the design of input. (Table 1)
- Observation 5: Overall the SciBERT models for both ADE/Reason-Drug is biased towards FP than TN, that is, positive samples are preferred more. (Table 1)
- Observation 6: For ADE-Drug, errors in relation extraction can be attributed to causes such as 'tagging inaccuracy' or 'abbreviation decoding' for True Negative case. (Table 2)
- Observation 7: For ADE-Drug, errors in relation extraction can be attributed to causes such as 'similar words' or 'lack of context' for False Positive case. (Table 2)
- Observation 8: For Reason-Drug, errors in relation extraction can be attributed to causes like 'error in positional tagging' or 'unfamiliar abbreviation' for True Negative case. (Table 3)
- Observation 9: For Reason-Drug, when a negative label is misclassified it can be due to 'lack of numerical knowledge' or 'decoding abbreviation' for False Positive. (Table 3)
- Observation 10: As opposed to Mid-Term, the dataset has more negative samples compared to positive, which may account for the change in recall score in Reason-Drug (Table 1)
- Observation 11: Reason Drug model has not performed as the ADE-Drug model, this can be due to the in-efficiency of the model in this task to effectively learn the embeddings of the newly added tokens and hence the relations between entities using the tokens.
- Observation 12: The macro and weighted average score has improved in both the models for positional tagging approach as compared to the mid-term. (Table 1)

Table 1: Classification Report

	ADE-Drug					
	Precision		Recall		F1-Score	
	Mid	Final	Mid	Final	Mid	Final
Positive Label	0.44	0.77	0.04	0.83	0.08	0.8
Negative label	0.51	0.78	0.95	0.72	0.67	0.75
Accuracy					0.51	0.78
Macro Avg	0.48	0.78	0.5	0.77	0.37	0.77
Weighted Avg	0.48	0.78	0.51	0.78	0.37	0.78

	Reason-Drug					
	Precision		Recall		F1-Score	
	Mid	Final	Mid	Final	Mid	Final
Positive Label	0.54	0.59	0.07	0.7	0.13	0.64
Negative label	0.52	0.57	0.94	0.45	0.67	0.55
Accuracy					0.52	0.58
Macro Avg	0.53	0.58	0.51	0.57	0.4	0.57
Weighted Avg	0.53	0.58	0.52	0.58	0.41	0.57

References:

Fig.1: https://medinform.jmir.org/2020/7/e18417/

Fig 2: Fei Li, Meishan Zhang, and Guohong A neural joint model for entity and relation extraction from biomedical text.

Table 2: Error Analysis ADE-Drug

ADE DRUG		
TEXT CORPUS	TYPE	POSSIBLE CAUSE
Due to concern for continued infection <DRUG>cefepime <DRUG> was added to .... <ADE>persistent fevers <ADE> on cefepime and	TN	inaccuracy in tagging entities in the sample
These medications were subsequently stopped by pcp due to <ADE> hypertension<ADE>Paranoid schizophrenia ... admit for psychosis after stopping <DRUG> risperidone<DRUG> for orthostatic hypotension	FP	Relation with similar words/tokens
Initial <ADE> AMS <ADE> was likely multifactorial due to ... medication induced from <DRUG> narcotics <DRUG> received vicodin	TN	Abbreviation decoding
Did just start <DRUG> iron <DRUG> supplementand <ADE> allergy <ADE> to a component is possible but time course	FP	Lack of context in medical literature

Table 3: Error Analysis Reason-Drug

REASON DRUG		
TEXT CORPUS	TYPE	POSSIBLE CAUSE
prefer to restart <DRUG> ASA</DRUG>P after any ... hemodynamically significant<REASON> GI bleed </REASON>ing	TN	Error in positional tagging ASA as ASA-P
She was given 1 unit <DRUG> FFP </DRUG> and 9 units PRBCs between ... It was decided after her second <REASON>hematoma </REASON> while on anticoagulation	TN	Unfamiliar with abbreviation
after her second hematoma while on<REASON>anticoagulation</REASON > the risks of ... she was not <DRUG>anticoagulated</DRUG>	FP	High correlation between word tokens
<DRUG> pain meds </DRUG> and POantibiotics ...e IV fluids <REASON> Pain control </REASON>	FP	Similar words