Clinical Judgement Study using Question Answering from Electronic Health Records

Bhanu Pratap Singh Rawat BRAWAT@UMASS.EDU *CICS Department*
University of Massachusetts
Amherst, MA, USA

Fei Li FEI_LI@UML.EDU *CS Department*
University of Massachusetts
Lowell, MA, USA

Hong Yu HONG_YU@UML.EDU *CS Department*
University of Massachusetts
Lowell, MA, USA

Abstract

Clinical judgement studies are essential for recognising the causal relation of a medication with adverse drug reactions (ADRs). Traditionally, these studies are conducted via expert manual chart review. By contrast, we propose an end-to-end deep learning question answering model to automatically infer such causal relations. Our proposed model identifies the causal relation by answering a subset of Naranjo questionnaire Naranjo et al. (1981) from electronic health records. It employs multi-level attention layers along with local and global context while answering these questions. Our proposed model achieves a macro-weighted F-score of $0.4598 - 0.5142$ across the selected questions and an overall F-score of 0.5011. We also did an ablation study to validate the importance of local and global context for the model.

1. Introduction

Pharmacovigilance and drug safety surveillance are quite sought after research subjects in the medical domain because of their importance for patience safety. Clinical judgement studies to extract causal relations between drugs and adverse drug reactions (ADRs) are essential parts of drug safety surveillance. An ADR can be loosely defined as any noxious, unintended or undesired effect of a medicine after doses used in humans for prophylaxis, diagnosis or therapy Naranjo et al. (1981). ADRs are the single largest contributor to hospital-related complications in inpatient settings Classen et al. (2011) and occur commonly at a rate of 2.4-5.2 per 100 hospitalized adult patients Bates et al. (1997); Classen et al. (1997); Nebeker et al. (2005). Anticoagulants are one of the most common drug classes that cause numerous ADRs, accounting for approximately 1 in every 10 of all drug-related adverse outcomes (specifically bleeding events) Lucado et al. (2006) and one-third of all ADRs among hospitalized Medicare patients Levinson and General (2010).

Technical Significance Prior research work in this domain has been mainly focused on extracting drug and ADR entities Jagannatha and Yu (2016); Aramaki et al. (2010) or identifying relations between them from electronic health records Munkhdalai et al. (2018).

However, inferring causality of ADRs using either linguistic cues or statistical models has significant limitations. For example, statistical correlation is very different from causality. Therefore, ADR relations detected by these methods still need to be validated by expert physicians. Also, clinical pharmacologists frequently disagree when analyzing the causality of ADRs Naranjo et al. (1981). As described by Naranjo et al. (1981), the suspected medication is usually confounded with other causes, and the adverse reaction cannot be easily distinguished from the manifestations of the disease making it significantly harder to extract the accurate relation between the drug and ADRs. Hence, there is a need of formulating the problem of causal relation extraction in a different way to automate such clinical studies.

Clinical Relevance Due to the lack of an established methodology for clinical studies, Naranjo scale was developed to standardize the causality assessment of ADRs Naranjo et al. (1981). Naranjo scale is frequently used by physicians to conduct causality assessment studies between a medication and ADRs Sharma et al. (2015); Shamna et al. (2014). It comprises of 10 questions and a subset of these questions is shown in Table 1. A causality scale (e.g., doubtful or probable) is assessed based on the answers to those questions. Naranjo scale has shown a marked improvement in within-raters agreement, reproducibility, reliability as compared to other approaches Naranjo et al. (1981). One strength of the Naranjo scale is that it can handle missing values: the scale is valid even with the answers to some of the questions. Therefore, Naranjo scale has been widely used as a standard in clinical domain.

Previous clinical judgement studies have solely been conducted on time-consuming manual chart reviews of electronic health records (EHRs) which require a lot of manual efforts put by experienced physicians. As such, these studies are usually conducted only on a subset of clinical notes due to multiple time constraints. To facilitate clinical judgement studies, we propose an end-to-end deep learning question answering model for automatically answering the Naranjo questionnaire by computing the causal scale as explained in section 2.1. Our model also utilizes the global and local context while answering the questions.

Our contributions are mainly three-fold:

1. We propose, an end-to-end deep learning model to answer the clinically-validated Naranjo questionnaire from EHRs to establish a causal relation between a medication and its ADRs. Our work is a huge contribution to drug safety surveillance and pharmacovigilance, as the current practice relies on the labour-intensive process of domain-experts who manually chart-review the EHRs.

2. By effectively integrating deep learning, multi-level attention and imbalanced learning, our model provides a decent macro-averaged f-score of 0.5011 across questions.

3. To the best of our knowledge, our model is the first deep learning model for clinical study question answering using EHRs. Our work could be used as a strong baseline for further related research.

#	Naranjo Questions	Yes	No	Do not know
1.	Are there previous conclusive reports on this reaction?	1	0	0
2.	Did the adverse event occur after the suspected drug was administered?	2	-1	0
3.	Did the adverse reaction improve when the drug was discontinued or a specific antagonist was administered?	1	0	0
4.	Did the adverse reaction reappear when the drug was readministered?	2	-1	0
5.	Are there alternative causes (other than the drug) that could have on their own cause the reaction?	-1	2	0
6.	Did the reaction reappear when a placebo was given?	-1	1	0
7.	Was the drug detected in the blood (or other fluids) in concentrations known to be toxic?	1	0	0
8.	Was the reaction more severe when the dose was increased or less severe when the dose was decreased?	1	0	0
9.	Did the patient have a similar reaction to the same or similar drugs in any previous exposure?	1	0	0
10.	Was the adverse event confirmed by any objective evidence?	1	0	0

Table 1: Naranjo Scale Questionnaire.

2. Naranjo Scale and Dataset

2.1. Naranjo Scale

The Naranjo Scale Questionnaire consists of 10 questions which are administered for each patient's clinical note. Each question can be answered as "Yes", "No" or "Do not know" , where "Do not know" is marked when the quality of the data does not allow an affirmative (yes) or negative (no) answer.

A score of $\{-1, 0, 1, 2\}$ is assigned to each question as shown in Table 1. The Naranjo scale assigns a causality score, which is the sum of the scores of all questions, that falls into one of four causality types: doubtful (≤ 0), possible ($1-4$), probable ($5-8$), and definite (≥ 9). In clinical settings, it is typically rare to find answers for all 10 Naranjo questions. The Naranjo scale is designed such that it is valid even if the answers for only a subset of the Naranjo questionnaire are provided.

2.2. Cohort Selection

We built an expert annotated EHR cohort to be used for training and evaluation of our proposed model. We selected the clinical notes of patients who were administered the anti-coagulant *Coumadin*. To increase the chance that the notes also contain ADRs, we focused on the patients who had any signs of internal bleeding such as gastrointestinal bleeding, blood clots or black tarry stools as these are the most common ADRs of anticoagulants. Physician annotators manually examined those notes and provided answers for each Naranjo question. The physicians provided granular information by annotating the relevant sentence in the EHR and then the answer of the related Naranjo question as one of the three answers: 'Yes', 'No' and 'Do not know'.

2.3. Dataset

Our dataset consists of discharge summaries of 446 unique patients. Since some of the patients were admitted more than once, there are 584 discharge summaries in total. Four physicians, supervised by a senior physician, annotated the Naranjo scale questionnaire for each of these discharge summaries. Each discharge summary was annotated by one of the four physician independently. Reconciliation was done by the senior physician who examined every annotation and discussed the differences with other physicians. Each discharge summary could have multiple ADRs, each of which could have a different Naranjo questionnaire. Our model attempts to detect all of the ADRs and their corresponding questionnaires and answers.

Since we are only interested in the questions that can be answered from the information provided in the discharge summary, we omitted the first question from our study. All the remaining questions were answered by the physicians. However, most of the answers (90% or more), for 4 questions, out of the remaining 9, were "Do not know". To build effective computational models with sufficient amount of data, we focused on the remaining 5 questions: 2, 3, 5, 7 and 10 as per Table 1. As described earlier, the imbalanced answer distribution is typical for Naranjo scale assessment and it would still be clinically meaningful even if only a subset of the Naranjo questions could be answered. The distribution of classes for these questions is given below in Table 2. The selected 5 questions still face the data imbalance challenge but they have a decent representation for each answer. For example, relevant sentences account for less than 5% of total sentences. Therefore, our end-to-end model integrates approaches to counter class imbalance, which is discussed in section 3.3.

Question #	Yes	No	Do not know
2	1633	139	666
3	381	21	181
5	2186	221	316
7	619	29	76
10	1683	678	227

Table 2: Distribution of answers for selected 5 questions.

3. Methodology

3.1. Problem Formulation

As mentioned in the previous section, a discharge summary can have multiple ADRs and their corresponding Naranjo questions. Our annotators went through each of the clinical note meticulously and annotated all the ADRs with their corresponding Naranjo question-answers. The annotation resulted in two levels of information: *relevant* sentence for which the Naranjo question has been answered and *answer* ("Yes", "No", and "Do not Know") for the specific Naranjo question. For example, the sentence "In ED, she was found to have a hgb of 9, INR 3.6, and rectal exam in ED revealed maroon stool" as shown in Figure 1 was annotated as a relevant sentence to answer the Naranjo question 2 for the ADR "maroon stool" (the answer is "yes").

According to the above explained formulation each of the *relevant* sentence would have one of the three answers ('Yes', 'No' and 'Do not know') for the Naranjo questions but the *non-relevant* sentences won't have any gold-standard training label. The model needs to learn to classify *non-relevant* sentences as well and hence, we labeled all the *non-relevant* sentences with a label: 'Non-relevant'. Thus, each sentence would have one of the four labels as answer to the selected Naranjo questions: 'Yes', 'No', 'Do not know' and 'Non-relevant'.

3.2. Clinical Question Answering Model

Our proposed model has three main parts: building sentence representation, adding local context to the sentence representation and using global context from the discharge summary.

3.2.1. Sentence representation

The sentence representation needs to be question-aware, hence, the question and the sentence are concatenated together and then sent as an input to the model. To build this part, we used a bidirectional long short-term memory (BiLSTM) network Graves and Schmidhuber (2005) with both self-attention Vaswani et al. (2017) and global attention (BiLSTM-Attn) Luong et al. (2015). The BiLSTM has 2 LSTM units where the first unit $\overrightarrow{LSTM}$ propagates in the forward direction and the second $\overleftarrow{LSTM}$ propagates in the backward direction. The hidden states from both LSTM units are concatenated to form the final hidden state.

$$\overrightarrow{LSTM}(x_t) + \overleftarrow{LSTM}(x_t) = \overrightarrow{h_t} + \overleftarrow{h_t} = h_t, \tag{1}$$

where $X \in [x_1, x_2, ...x_n]$ denotes the concatenated question and sentence tokens, h_t indicates the hidden state at the time step t. These hidden representations $H = [h_1, h_2....h_t...h_n]$ are then passed through a self-attention layer that takes these hidden states as input in the form of three matrices: *query (Q)*, *key (K)* and *value (V)*. For self-attention, these three matrices are equal to each other and hence, H is passed as Q, K and V to the self-attention layer.

$$H' = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

where Q, K, $V = H$ and d_k is the dimension of key which is the dimension of hidden state in self-attention. $H' = [h'_1, h'_2....h'_t...h'_n]$ are the new hidden states after self-attention. These hidden states (H') are then passed through an affine layer (W_a) and a softmax layer to get the location-based global attention Luong et al. (2015), given by:

$$a_t = softmax(W_a h'_t), \tag{3}$$

where a_t denotes the weight of each hidden state and the final hidden representation H'_f can be obtained by the weighted sum operation:

$$H'_f = a_t h'_t. \tag{4}$$

Self-attention helps in building *question-aware* representation of the tokens and global attention helps in getting the final sentence representation over these question-aware token representation.

3.2.2. Local Context Representation

Along with the sentence, we used contextual information in the form of neighboring sentences which are present before and after the annotated sentence. We tuned the context window as a hyper-parameter. Each context sentence is passed through the *BiLSTM-Attn* model as explained in section 3.2.1 to get the final hidden representation (H'_f) of each sentence.

A context window of 3 corresponds to 7 final hidden representations: $[H'_{f,-3}, H'_{f,-2}, H'_{f,-1}, H'_{f,0}, H'_{f,1}, H'_{f,2}, H'_{f,3}]$. These hidden representations are passed through another BiLSTM model with global attention. This provides us the final representation at the context level (H'_{sent}) which is used to predict the answer for the Naranjo question.

3.2.3. Document Representation

In addition to sentence representations we also added document representation, which provides global information while answering the question. The entire discharge summary is tokenized and concatenated to the tokens of the question for which it is being used. The concatenated vector is passed through a *BiLSTM-Attn* network to get the *question-aware* document representation. The representation of the discharge summary H'_{doc} is concatenated with the representation of the sentence and its context to get the vector $H'_{answer} = [H'_{sent}, H'_{doc}]$. The concatenated vector (H'_{answer}) is then passed through the inference layer to predict the answer for the Naranjo question: *Yes*, *No*, *Do not know* or *Non-relevant*.

Our proposed clinical question answering model is illustrated in Fig. 1.

3.2.4. Inference

If the Naranjo answer for the sentence is predicted as 'Yes', 'No' or 'Do not know', it is also recognized as *Relevant*. If the label of the sentence is predicted 'Non-relevant', it is recognized as *Non-Relevant*. Therefore, for *inference* the final answer predicted by the model out of the four labels is enough.

3.3. Approaches for Data Imbalance

We observe in Table 2 that relevant sentences are only a small fraction of the sentences in an EHR and accordingly the answers to Naranjo questions are quite imbalanced. Take question 2 as an example, the ratio of the labels is Non-relevant : Yes : No : Do not know= 36799 : 1299 : 124 : 547 because we add another label (*Non-relevant*) for all the sentences which didn't have an annotated answer. This poses a challenge for our model as this unbalancing is quite significant. In order to tackle this problem, we integrate two techniques.

Weighted Loss: We implemented weighted loss technique Zhou and Liu (2006) where the total loss is calculated as weighted sum of loss according to the class. Log weighing helps in smoothing the weights for highly unbalanced classes, which is the case in our dataset as shown in Table 2

$$w_{c,q} = \begin{cases} 1.0 & \text{if}(w_{c,q} < 1.0) \\ log(\alpha * T_q / T_{c,q}) & \end{cases} \tag{5}$$

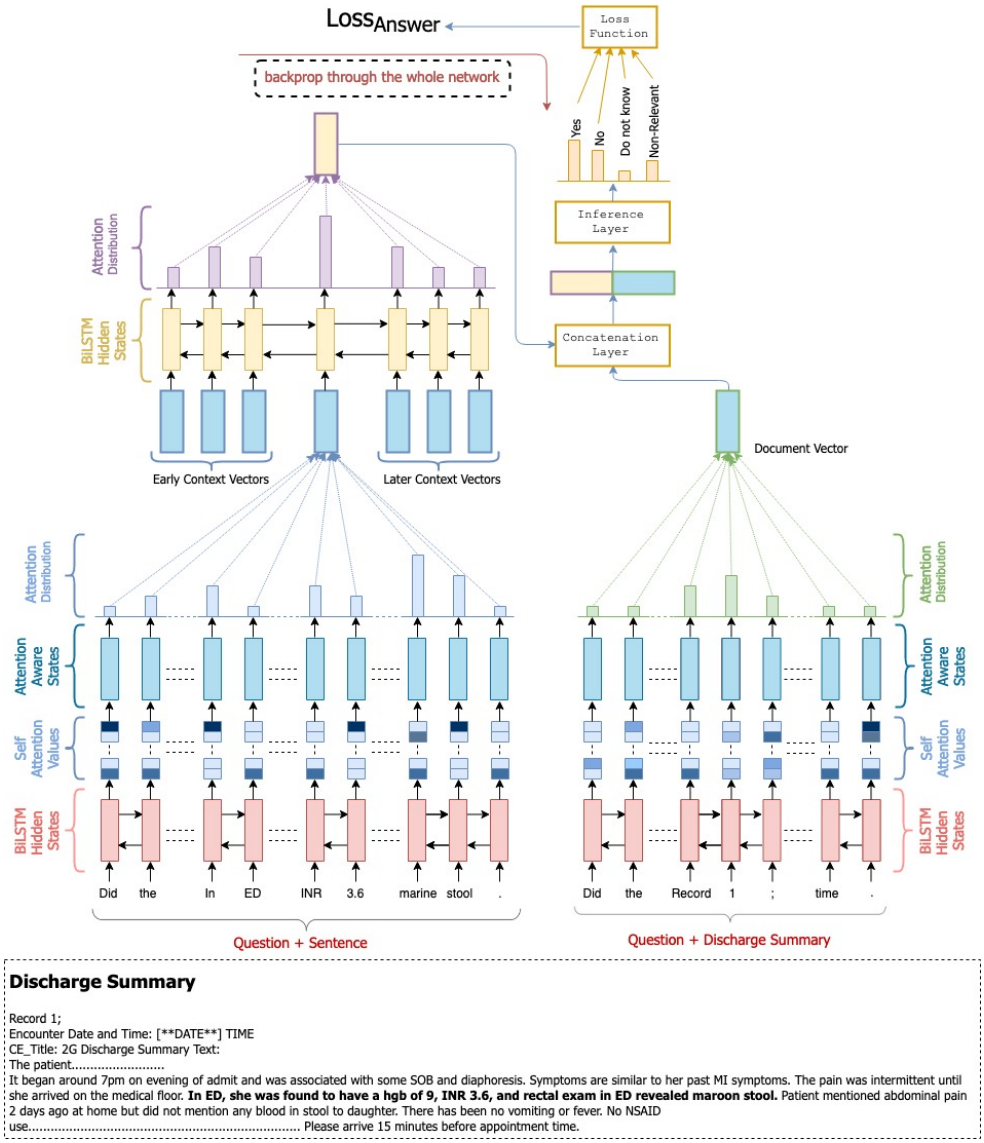

Figure 1: Clinical Question Answering Model. The lower-level attention layers (blue and green) attend to the words in a sentence or a document. The higher-level attention layer (purple) attends to the sentence and its neighbouring sentences which act as the *local* context. The document vector is created separately which act as the *global* context.

Where $q \in \{2, 3, 5, 7, 10\}$ represents one of the 5 questions and $c \in \{\text{Yes, No, Do not know, Non-relevant}\}$ and we tuned α as a hyperparameter. T_q is the count for question q and $T_{c,q}$ is the count of a particular class c for question q.

Down-sampling: We down-sampled the *Non-relevant* label sentences. Before each training epoch, we randomly reduced the training data for *Non-relevant* label by $X\%$. We tuned X as another hyperparameter for the model.

		Weighted			Macro			Micro		
Ques #	**Model Config**	**W-P**	**W-R**	**W-F**	**MA-P**	**MA-R**	**MA-F**	**MI-P**	**MI-R**	**MI-F**
Ques 2	**NDNC**	0.9688	0.9468	0.9566	0.3902	0.4890	0.4196	0.9468	0.9468	0.9468
	NDC	0.9677	0.9477	0.9562	0.4374	0.5457	0.4738	0.9477	0.9477	0.9477
	DC	0.9659	0.9507	**0.9570**	0.4500	0.5841	**0.4919**	0.9507	0.9507	**0.9507**
Ques 3	**NDNC**	0.9887	0.9829	0.9857	0.3168	0.3554	0.3299	0.9829	0.9829	0.9829
	NDC	0.9891	0.9767	0.9825	0.3967	0.5054	0.4244	0.9767	0.9767	0.9767
	DC	0.9889	0.9892	**0.9890**	0.5180	0.5139	**0.5142**	0.9892	0.9892	**0.9892**
Ques 5	**NDNC**	0.9560	0.9394	0.9463	0.4217	0.4801	0.4408	0.9394	0.9394	0.9394
	NDC	0.9612	0.9413	0.9498	0.4117	0.5261	0.4515	0.9413	0.9413	0.9413
	DC	0.9611	0.9454	**0.9520**	0.4349	0.5271	**0.4681**	0.9454	0.9454	**0.9454**
Ques 7	**NDNC**	0.9914	0.9886	**0.9898**	0.4103	0.4979	0.4437	0.9886	0.9886	**0.9886**
	NDC	0.9912	0.9853	0.9876	0.3814	0.5000	0.4216	0.9853	0.9853	0.9853
	DC	0.9899	0.9878	0.9886	0.4524	0.4829	**0.4598**	0.9878	0.9878	0.9878
Ques 10	**NDNC**	0.9604	0.9292	0.9420	0.4085	0.5804	0.4608	0.9292	0.9292	0.9292
	NDC	0.9617	0.9353	**0.9457**	0.4307	0.5776	0.4769	0.9353	0.9353	**0.9353**
	DC	0.9583	0.9349	0.9445	0.4498	0.5815	**0.4917**	0.9349	0.9349	0.9349
Overall	**NDNC**	0.9741	0.9578	0.9649	0.4247	0.5714	0.4759	0.9578	0.9578	0.9578
	NDC	0.9741	0.9581	0.9648	0.4378	0.5777	0.4867	0.9581	0.9581	0.9581
	DC	0.9719	0.9604	**0.9652**	0.4617	0.5725	**0.5011**	0.9604	0.9604	**0.9604**

Table 3: Weighted, macro-averaged and micro-averaged Precision, Recall and F-score values for all models. Best performance F-scores have been highlighted above. NDNC: neither document (global) nor local context representation is provided as input to the model. NDC: No document representation but local context is provided, DC: both document and local context representation is provided as input.

4. Experiments and Results

4.1. Experimental Settings

For each model, the dataset was divided into three sets such that each question has an equal distribution of train : validation : test = 60 : 20 : 20. We used pre-trained embeddings learned from Pubmed abstracts, DrugBank descriptions, DailyMed prescriptions and about two thousand clinical notes, available to us, via fasttext Bojanowski et al. (2016) technique. Along with network hyper-parameters, we tuned percentage down-sampling as well as log-weighing parameter (α in eqn 5) as hyperparameters. The values of hyperparameters for each model have been provided in Appendix A.

4.2. Evaluation metrics

We utilize weighted, macro-averaged and micro-averaged precision recall and f-score for the model across questions in Table 3. Micro-average reports the average of instance-level performance and therefore is biased to the label with the highest frequency count which is why the micro-averaged metric values are quite high for all models as they are biased towards the *Non-relevant* label. Similarly, weighted metrics are also biased towards labels with higher frequency and are quite high for all models. Macro-averaged metrics are

calculated by averaging the performance across the labels and thus provide better insight on model's performance across different labels.

4.3. Results

In this sub-section, we would discuss some of the results provided in Table 3. We would mainly look at the performance of our proposed model, referred to as *DC* model configuration, in terms of precision, recall and f-score with respect to all the selected questions. As has been mentioned in section 4.2, weighted and micro-averaged metrics are biased towards the label but are essential as they can be used to compare the overall performance of the model whereas macro-averaged metrics help in comparing the performance of models across different labels while considering the imbalance in the data.

Our model (DC configuration) achieved an overall macro f-score of 0.5011 while maintaining comparable precision and recall values. It also achieved an overall micro-averaged f-score of 0.9604 and weighted average f-score of 0.9652. This shows that model has a good overall performance and has good performance across the four labels: 'Yes', 'No', 'Do not know' and 'Non-relevant'. For three out of the selected five questions, the macro-weighted f-score is close to 0.50 which shows that along with performing well across labels, the model performs significantly well across different questions. This suggests that it is able to learn the generalized information and generate *question aware* sentence and document representations.

4.4. Ablation Study

In order to understand the effectiveness of the global and local context provided to the model in the form of document representation and neighboring sentence representations, we also performed an ablation study where we developed two variants of our model with different configurations. The first variant of the model is only provided with the sentence as an input without global or local context and is referred by *NDNC* configuration in Table 3. The second variant of the model is provided with the local context in the form of neighboring sentence representations but not global context and has been referred to as *NDC* configuration in Table 3. The weighted and micro-weighted performance values are super close to each other and hence it's hard to draw any conclusive inferences from them regarding the performance improvement of the model, hence, we would mainly focus on the macro-weighted evaluation metrics.

For macro-weighted f-score, across all the questions our model performs better than the two variants of the model. This shows that global and local context do provide extra information to the model in order to make accurate answer predictions. For question 2 and 3, we can see that there is major improvement in the performance of our model when local context is provided to it as an input. Macro-weighted f-score improved by ~ 0.1 for both question 2 and 3. In question 5 and 10, there is a small improvement in the performance of the model, resulting in an increment of ~ 0.01 in the macro-weighted f-score. Only in question 7, the performance of the model decreased slightly when the local context is added to it but since there is an overall performance improvement and improvement in four out of five questions, addition of the local context to the model is quite justified.

It can be observed in Table 3 that there is consistent improvement in the performance of our model when document representation is provided as input. Specifically in question 3, there is an improvement of ~ 0.1 in the macro-weighted F-score of the model when document representation is provided as input. The consistent improvement in the performance of our model with document representation definitely suggests that global context is an important part of the model to predict accurate answers for these questions. On observing the overall macro-weighted precision in Table 3, we can see that the precision of the model consistently improves with the addition of local and global context to its architecture. There is a significant improvement in the overall macro-weighted precision of the model which mainly contributes in the improvement of the overall f-score, the overall macro-weighted recall also improves but not significantly.

5. Related Works

Multiple studies have been conducted using the Naranjo Scale Naranjo et al. (1981) to identify the causal relationship between an ADR and a medication Davies et al. (2006); Arulmani et al. (2008). Most of these clinical judgement studies have been conducted on manual chart reviews Davies et al. (2006); Priyadharsini et al. (2011); Passarelli et al. (2005) whereas only some of them performed statistical analysis of the relationships between drug and ADRs Davies et al. (2009) which was also done manually.

There have been efforts on identifying the ADRs and medications using different statistical methods Huynh et al. (2016); Harpaz et al. (2010); Jagannatha and Yu (2016). However, these methods did not focus on extracting relations between them. Further studies were conducted to extract the relationship between ADRs and medications Munkhdalai et al. (2018); Feldman et al. (2015); Sondhi et al. (2012), but focused only on extracting the existence of a relation. They are not able to answer the causal relationship between a pair of ADR and medication. Moreover, these studies did not use the global information provided in the clinical note.

In order to utilize both powerful deep learning techniques and information provided by the Naranjo Scale, we conducted our research on developing a *question answering* model. Enormous efforts have been put in to develop effective question answering techniques Min et al. (2017); Wang et al. (2017); Seo et al. (2016). Seo et al. Seo et al. (2016) designed a bi-directional attention model to make full use of query and context information. Although we also leveraged the attention method, our task is essentially different from extractive question answering since our task needs models to perform various inferences related to the Naranjo Scale Questionnaire. Thus, we developed a deep learning model which utilises self-attention Vaswani et al. (2017) and global attention Luong et al. (2015) to build question aware sentence representations which can be used to predict the answers for Naranjo questions. As we observed in section 4.3 that our proposed model generalizes well over different questions suggesting that it is able to learn question-aware representations and patterns. We also observed in section 4.4 that adding local and global context in the form of neighboring sentence and discharge summary representation helps in improving the performance of the base model. The global and local context provides the model additional information to make accurate predictions.

6. Conclusion

In this paper we proposed an end-to-end deep learning model which integrates self attention Vaswani et al. (2017) and global attention Luong et al. (2015), which can automatically answer a subset of Naranjo Questionnaire Naranjo et al. (1981). We chose Naranjo questionnaire as it is well accepted by physicians for assessing the causality between a medication and its ADRs. We show the importance of global and local context for predicting the answers for these questions with the help of ablation study discussed in section 4.4. Our model achieves an overall macro-weighted f-score of 0.5011 and generalizes well across the subset of these questions. Moreover, we employed down-sampling and weighted loss which assist our model in dealing with the problem of high class-imbalance, which is quite frequent in medical domain. To the best of our knowledge, our study is the first of its kind to automate the process of clinical judgement study with the help of question answering approach. Our work can facilitate several research domains such as pharmacovigilance, biomedical NLP and clinical decision support and can also be used as a strong baseline[1] for future work in this direction.

References

Eiji Aramaki, Yasuhide Miura, Masatsugu Tonoike, Tomoko Ohkuma, Hiroshi Masuichi, Kayo Waki, and Kazuhiko Ohe. Extraction of adverse drug effects from clinical records. *MedInfo*, 160:739–743, 2010.

R Arulmani, SD Rajendran, and B Suresh. Adverse drug reaction monitoring in a secondary care hospital in south india. *British journal of clinical pharmacology*, 65(2):210–216, 2008.

David W Bates, Nathan Spell, David J Cullen, Elisabeth Burdick, Nan Laird, Laura A Petersen, Stephen D Small, Bobbie J Sweitzer, and Lucian L Leape. The costs of adverse drug events in hospitalized patients. *Jama*, 277(4):307–311, 1997.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

David C Classen, Stanley L Pestotnik, R Scott Evans, James F Lloyd, and John P Burke. Adverse drug events in hospitalized patients: excess length of stay, extra costs, and attributable mortality. *Jama*, 277(4):301–306, 1997.

David C Classen, Roger Resar, Frances Griffin, Frank Federico, Terri Frankel, Nancy Kimmel, John C Whittington, Allan Frankel, Andrew Seger, and Brent C James. global trigger toolshows that adverse events in hospitals may be ten times greater than previously measured. *Health affairs*, 30(4):581–589, 2011.

EC Davies, CF Green, DR Mottram, and M Pirmohamed. Adverse drug reactions in hospital in-patients: a pilot study. *Journal of clinical pharmacy and therapeutics*, 31(4): 335–341, 2006.

1. https://github.com/bsinghpratap/clinicalQA_mlhc

Emma C Davies, Christopher F Green, Stephen Taylor, Paula R Williamson, David R Mottram, and Munir Pirmohamed. Adverse drug reactions in hospital in-patients: a prospective analysis of 3695 patient-episodes. *PLoS one*, 4(2):e4439, 2009.

Ronen Feldman, Oded Netzer, Aviv Peretz, and Binyamin Rosenfeld. Utilizing text mining on online medical forums to predict label change due to adverse drug reactions. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1779–1788. ACM, 2015.

Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005.

Rave Harpaz, Krystl Haerian, Herbert S Chase, and Carol Friedman. Mining electronic health records for adverse drug effects using regression based methods. In *Proceedings of the 1st ACM International Health Informatics Symposium*, pages 100–107. ACM, 2010.

Trung Huynh, Yulan He, Alistair Willis, and Stefan Rüger. Adverse drug reaction classification with deep neural networks. Coling, 2016.

Abhyuday N Jagannatha and Hong Yu. Bidirectional rnn for medical event detection in electronic health records. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2016, page 473. NIH Public Access, 2016.

Daniel R Levinson and Inspector General. Adverse events in hospitals: national incidence among medicare beneficiaries. *Department of Health and Human Services Office of the Inspector General*, 2010.

Jennifer Lucado, Kathryn Paez, and A Elixhauser. Medication-related adverse outcomes in us hospitals and emergency departments, 2008: statistical brief# 109. 2006.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

Sewon Min, Minjoon Seo, and Hannaneh Hajishirzi. Question answering through transfer learning from large fine-grained supervision data. *arXiv preprint arXiv:1702.02171*, 2017.

Tsendsuren Munkhdalai, Feifan Liu, and Hong Yu. Clinical relation extraction toward drug safety surveillance using electronic health record narratives: Classical learning versus deep learning. *JMIR public health and surveillance*, 4(2), 2018.

Cláudio A Naranjo, Usoa Busto, Edward M Sellers, P Sandor, I Ruiz, EA Roberts, E Janecek, C Domecq, and DJ Greenblatt. A method for estimating the probability of adverse drug reactions. *Clinical Pharmacology & Therapeutics*, 30(2):239–245, 1981.

Jonathan R Nebeker, Jennifer M Hoffman, Charlene R Weir, Charles L Bennett, and John F Hurdle. High rates of adverse drug events in a highly computerized hospital. *Archives of internal medicine*, 165(10):1111–1116, 2005.

Maria Cristina G Passarelli, Wilson Jacob-Filho, and Albert Figueras. Adverse drug reactions in an elderly hospitalised population. *Drugs & aging*, 22(9):767–777, 2005.

R Priyadharsini, A Surendiran, C Adithan, S Sreenivasan, and Firoj Kumar Sahoo. A study of adverse drug reactions in pediatric patients. *Journal of pharmacology & pharmacotherapeutics*, 2(4):277, 2011.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*, 2016.

M Shamna, C Dilip, M Ajmal, P Linu Mohan, C Shinu, CP Jafer, and Yahiya Mohammed. A prospective study on adverse drug reactions of antibiotics in a tertiary care hospital. *Saudi pharmaceutical journal*, 22(4):303–308, 2014.

Rohini Sharma, Devraj Dogra, and Naina Dogra. A study of cutaneous adverse drug reactions at a tertiary center in jammu, india. *Indian dermatology online journal*, 6(3): 168, 2015.

Parikshit Sondhi, Jimeng Sun, Hanghang Tong, and ChengXiang Zhai. Sympgraph: a framework for mining clinical notes through symptom relation graphs. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1167–1175. ACM, 2012.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 189–198, 2017.

Zhi-Hua Zhou and Xu-Ying Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, 18(1):63–77, 2006.

Appendix A.

Model Config	**BiLSTM Token Hidden Dim**	**BiLSTM Sent Hidden Dim**	**Embedding Dim**	**Learning Rate**	α	**Down-sampling %**
NDNC	500	460	200	0.0009	15	60
NDC	500	450	200	0.001	15	60
DC	500	450	200	0.001	15	60

Table 4: Hyper-parameter values for all models.