Naranjo Question Answering using End-to-End Multi-task Learning Model

Bhanu Pratap Singh Rawat
UMass Amherst
Amherst, USA
brawat@cs.umass.edu

Fei Li
UMass Lowell
Lowell, USA
fei_li@uml.edu

Hong Yu
UMass Lowell
Lowell, USA
hong.yu@umassmed.edu

ABSTRACT

In the clinical domain, it is important to understand whether an adverse drug reaction (ADR) is caused by a particular medication. Clinical judgement studies help judge the causal relation between a medication and its ADRs. In this study, we present the first attempt to automatically infer the causality between a drug and an ADR from electronic health records (EHRs) by answering the Naranjo questionnaire, the validated clinical question answering set used by domain experts for ADR causality assessment. Using physicians' annotation as the gold standard, our proposed joint model, which uses multi-task learning to predict the answers of a subset of the Naranjo questionnaire, significantly outperforms the baseline pipeline model with a good margin, achieving a macro-weighted f-score between $0.3652 - 0.5271$ and micro-weighted f-score between $0.9523 - 0.9918$.

KEYWORDS

Question answering, RNN, LSTM, Naranjo questionnaire, Multi-task learning, Attention based network

ACM Reference Format:
Bhanu Pratap Singh Rawat, Fei Li, and Hong Yu. 2019. Naranjo Question Answering using End-to-End Multi-task Learning Model. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3292500.3330770

1 INTRODUCTION

Causal inference remains an unsolved task in statistical and machine learning. In the clinical domain, identifying the causality between a medication and its adverse drug reaction (ADR) is essential for pharmacovigilance and drug safety surveillance. An ADR is defined as any noxious, unintended and undesired effect of a drug after doses used in humans for prophylaxis, diagnosis, or therapy [22]. ADRs are the single largest contributor to hospital-related complications in inpatient settings [7]. ADRs affect more than two million hospital stays annually [16], and prolong hospital length of stay by 1.7 to 4.6 days [2, 6]. Anticoagulants are among the most commonly implicated drug classes in ADRs, accounting for approximately 1 in every 10 of all drug-related adverse outcomes (e.g., bleeding) in hospital settings [18].

However, identifying the causality between a medication and its ADRs is challenging. Clinical pharmacologists frequently disagree when analyzing the causality of ADRs [22]. As described by [22], the challenges include that manifestations of ADRs are nonspecific, that the suspected medication is usually confounded with other causes, and that the adverse reaction often cannot be distinguished from manifestations of the disease.

As a result, the Naranjo scale was developed to standardize assessment of causality for all adverse drug reactions [22]. The Naranjo scale comprises a list of 10 questions (Table 1). A causality scale (e.g., probable) is then assessed based on the answers to those questions. Naranjo scale has showed a marked improvement in between-raters and within-raters agreement while assessing ADRs as compared to other approaches. The reproducibility of the scale has been maintained on retesting along with a high intra-class correlation coefficient of reliability [22]. Therefore, Naranjo scale has been widely used as a standard since its inception.

Previous clinical judgement studies have solely been conducted on manual chart reviews of electronic health records (EHRs). In this work, we have developed natural language processing (NLP) techniques to for automated Naranjo question answering assessment. Researchers have developed different NLP approaches for electronic health records (EHRs) applications, including medical entity extraction [15], and relation identification [17]. Previously, ADRs have been detected based on context or linguistic cues (e.g., "caused by") indicative of the causal relations between a medication and ADRs [21]. There have also been studies to identify ADRs [13, 14] using statistical models [12]. However, inferring the causality of ADRs using either linguistic cues or statistical models have significant limitations. For example, physicians may not explicitly describe the causality between a medication and its ADEs, and statistical correlation is far from causality. Therefore, the ADRs detected by both methods still need to be clinically validated.

In this study, we propose a joint model (JM) for automatically answering the Naranjo questionnaire by computing the causality scale as explained in Section 3.1. The joint model utilizes multi-task learning to identify relevant sentences and learns from their answers for different questions. To evaluate our model, we evaluated whether a widely used anticoagulant, Coumadin, causes an ADR using longitudinal EHRs. Coumadin is a commonly prescribed drug for patients with cardiovascular diseases. It can have interim side effects such as nausea or loss of appetite as well as serious side effects such as internal bleeding, which can continue for up to a week after discontinuing the medication.

KDD '19, August 4–8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3330770

To build the joint model (JM), we first built an annotated EHR cohort to be used for training and evaluation. We selected the clinical notes (mainly discharge summaries) of patients who were administered Coumadin. To increase the chance that the notes also contain ADRs, we focused on 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 highly granular information by annotating the *relevant* text in the EHR and then the answer of the related Naranjo question as one of the three answers: 'Yes', 'No' and 'Do not know'.

This provides us two important information: the *relevant* sentence and *answer* for each question in the questionnaire, although not all questions are answered in EHRs because there might be no relevant or conclusive information to answer them. The questionnaire and the dataset are described in more detail in Section 3. The problem formulation for the multi-task learning has been provided in Section 4.1 and our methodology is described in Section 4 followed by results and analysis of models in Section 5. We found that our proposed model outperformed the baseline Pipeline model for 4 out of 5 questions and achieves macro-averaged f-score in the range of $0.3652 - 0.5271$ for all selected questions.
Our contributions are mainly three fold:

(1) We infer causality between a medication and its adverse drug reaction(s) by building a computational model to automatically answer the clinically validated Naranjo questionaire using EHRs. Our work is a significant contribution to patient 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) We formulate this Naranjo questionnaire as a novel end-to-end multi-task learning problem. Our multi-task joint model uses multi-attentions, enriched with contextual information. Our joint model also integrates cost-sensitive learning and down sampling to mitigate the data imbalance challenge.
(3) To the best of our knowledge, our work is the first computation model for clinical question answering using EHRs. Our model could be used as a strong baseline for any further research work in this area.

2 RELATED WORKS

Multiple studies have been conducted using the Naranjo Scale [22] to identify the causal relationship between an ADR and a medication [1, 8]. Most of these clinical judgment studies have been conducted using manual chart reviews [8, 23, 25] whereas only some of them performed statistical analysis on the relationships between medications and adverse drug reactions [9]. The statistical analysis was also performed manually after extracting the dosage of the medications and occurrence of ADRs.

There have been efforts on identifying the ADRs and medications using different statistical methods [12, 14, 15]. However, these methods did not identify the causal relations between them. There have been further studies to extract the relations between ADRs and medications [10, 21, 27], but still they just focused on extracting if there is a relation between a pair of entities, and failed to answer the causality. Moreover, these studies focused on local context and did not use the information from the whole clinical note.

In order to utilize both the powerful deep learning techniques and information provided by the Naranjo Scale, we built a *question answering* model. Enormous efforts have been put in to develop effective question answering techniques [20, 26, 28]. Seo et al. [26] 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.

Our work is also related to multi-task learning [5], since we designed a joint model to perform relevant sentence detection and answer prediction simultaneously. There have been efforts on training question answering models via multi-task learning as well [3, 4]. These studies leveraged evidences and answers, which are both provided within the text. They also utilized queries to find the evidences in the context. However, in our case the questions are fixed, which makes it more difficult to train models. Furthermore, we did not only utilize the evidence of the answer, but also recognize whether a sentence is relevant to the answer of our Naranjo questions. Hence, we focused on developing a *question answering* technique which can answer the Naranjo questions for each clinical record by focusing on two objectives: classifying relevant sentences and predicting correct answers. Thereafter, our model could be utilized to calculate the Naranjo causal score as explained in section 3.1.

3 NARANJO SCALE AND DATASET

3.1 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.

Table 1: Naranjo Scale Questionnaire.

#	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

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 Naranjo questions, that falls into one of four causality types: doubtful (≤ 0), possible ($1-4$), probable ($5-8$),

and definite ($\geq$ 9). In clinical settings, it is typically rare to find answers to all 10 Naranjo questions. The Naranjo scale is designed so that it is valid even if the answers for only a subset of the Naranjo questionnaire can be found.

3.2 Dataset

Our EHR dataset consists of the discharge summaries of 446 patients with cardiovascular diseases. Since some of the patients were admitted more than once, there are 584 discharge summaries in total. Four trained annotators, 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 annotators independently. Reconciliation was done by a senior physician who examined every annotation and discussed the difference with other physicians. Each discharge summary could have multiple ADRs, each of which could have multiple Naranjo questionnaires. Our joint model attempts to detect all of the ADRs and the corresponding questionaires and answers.

Since we are only interested in the questions that can be answered within a discharge summary, we omitted question 1 from our study. For the remaining questions, most of the answers, 90% or more, for questions 4, 6, 8 and 9 were "Do not know". As described earlier, the imbalanced answer distribution is typical for the Naranjo scale assessment and it would still be clinically meaningful even if only a subset of the Naranjo questions could be answered. To build effective computational models with sufficient amount of data, we focused on the remaining 5 questions: 2, 3, 5, 7 and 10. The distribution of classes for these questions is given below in Table 2. For each model, the training, validation and test split was 60 : 20 : 20. In the selected 5 questions, we tried to account for class imbalance during modeling which is discussed in Section 4.5.

Table 2: Distribution of answers for selected 5 questions.

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

4 METHODOLOGY

4.1 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 types 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").

In summary, we have the answers of Naranjo questions as well as the sentences which contain the relevant information for answering the questions. The tasks are summarized as follow:

(1) *Classifying Relevant Sentences*: Each sentence from the discharge summary would be passed to the *relevance* model which would classify the sentence as: *Relevant* or *Non-Relevant*.
(2) *Predicting Answers*: The *relevant* sentences are then passed to the *question answering (QA)* model which would predict the answer of the Naranjo question as: "Yes", "No" or "Do not know".

We developed both a pipeline model as well as a multi-task joint model for every Naranjo question. Building a pipeline model is straight-forward but building an end-to-end joint model (JM) is more challenging. For example, if the *sentence relevant* model makes a mistake during the training (for example, classifying a *Non-Relevant* sentence as *Relevant*), the *QA* model would have no *answer* label to train on. To deal with this problem, we added an extra label for the *QA* model as *False*. If a statement is *Non-Relevant*, its gold-standard label for the *QA* model would be set to *False*. Therefore, the *QA* model predicts one of the four answers ("Yes", "No", "Do no know" or "False") for each statement.

4.2 Baseline Pipeline Model

For the pipeline model, we built two different sub-models for classifying relevant sentences and for predicting answers, respectively. Each model is explained in details below:

4.2.1 Relevance Model. The *relevance* model is a binary-classification model. To build this model, we used the bidirectional long short-term memory network [11] with global attention (BiLSTM-Attn) [19] over the tokens. 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 a sentence and its 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 an affine layer (W_a) and a softmax layer to get the location-based global attention [19], given by:

$$a_t = softmax(W_a h_t), \tag{2}$$

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

$$h' = a_t h_t. \tag{3}$$

This final hidden representation (h') is used to predict the *relevance* for the sentence. An illustration of the *relevance* model is shown in Fig. 1.

4.2.2 Question Answering Model (QA). The *QA* model is trained on only relevant sentences which have one of the three labels: "Yes", "No" or "Do not know". We evaluated whether the contextual information (meaning the sentences before or after the annotated relevant sentences) could be useful or not by adding the before and after sentences in our model. We tuned the context window

as a hyper-parameter. An illustration of the *QA* model is shown in Fig. 1. Each sentence is passed through a BiLSTM-Attn model as explained earlier to get the final hidden representation (h').

If we have a context window of 3, we would have 7 final hidden representations: $[h'_{-3}, h'_{-2}, h'_{-1}, h'_0, h'_1, h'_2, h'_3]$. These hidden representations are then passed through another BiLSTM-Attn model. After that, we get the final representation at the context level (h'_{sent}) which is used to predict the answer for the Naranjo question.

4.2.3 Inference. Each sentence first goes through the *relevance* model. If the model predicts that the sentence is *Non-relevant*, it becomes the final predicted label for the sentence. If the model predicts that the sentence is *Relevant*, the sentence along with its context sentences pass through the *QA* model. The answer predicted by the *QA* model becomes the final predicted label for that sentence.

4.3 Joint Model (JM)

As shown in Fig. 1, the joint model (JM) also has two sub-models, same as the ones explained in the last section. The main differences between the JM and pipeline models are as follow:

(1) JM is trained using *multi-task learning* [5] where both the objectives of classifying the *relevance* of the sentence and the Naranjo *answer* for that sentence are considered simultaneously. This is implemented by adding a virtual gold label *False* for *Non-relevant* sentence, as explained in Section 4.1. During training, the loss function of the joint model ($Loss_{JointModel}$) is formalized as:

$$Loss_{JointModel} = Loss_{Relevance} + Loss_{Answer} \tag{4}$$

Both $Loss_{Relevance}$ and $Loss_{Answer}$ are negative log likelihood losses [24] for their respective label predictions.

(2) The BiLSTM layer in JM are shared by the *relevance* and *QA* sub-models, as shown in Fig. 1. Concretely, the same BiLSTM units are used for getting the hidden outputs from the tokens of the sentence. The two submodels share the BiLSTM units as well as the parameters of the attention layer. Such a "shared model" is reasonable because both sub-models use similar sentence context for the classification. A separate BiLSTM attention layer is used for the final representations of the sentence and its' context sentences. Here the attention layer is different because the attention is over sentences and not tokens now.

4.3.1 Inference. If the Naranjo answer for the sentence is predicted as *Yes*, *No* or *Do not know* by the QA sub-model of JM, it is recognized as *relevant*. If the label of the sentence is *False*, it is recognized as *None-Relevant*. Therefore, only the answer predicted by the QA sub-model is necessary for evaluation.

4.4 Joint Model with Doc Representation (JM-Doc)

JM utilizes the information from the sentence and its surrounding context to predict the answer of a Naranjo question. In addition to the surrounding context, we believe that the global information from the entire discharge summary (doc) could be useful for the Naranjo questionnaire. In order to accommodate the document information, we added a BiLSTM-Attn sub-model to get the document representation, as shown in Fig. 1.

4.4.1 Document Representation. The entire discharge summary is first tokenized and then passed through a BiLSTM-Attn network. This BiLSTM network shares its parameters with the BiLSTM network used to build the representation for the sentences. The sentence-level BiLSTM-Attn network remains the same but we used another attention layer to generate the document representation. This document-level attention layer can learn how to transfer important global information to the further layers.

The final representation of the discharge summary h'_{doc} is then concatenated with the final 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 sentence: *Yes*, *No*, *Do not know* or *False*. The sentence and its context representation represent the local information while the representation of the whole discharge summary represent the global information. Our results show that both help answer a specific Naranjo question. The inference for this model is same as JM.

4.5 Class Imbalance

We can observe in Table 2 that the answers to Naranjo questions are quite unbalanced. This becomes an even bigger problem when we try to learn the joint model (JM). Take question 2 as an example, the ratio of labels for the Pipeline model is Yes : No : Do not know = 1299 : 124 : 547, the ratio for JM, however, changes to False : Yes : No : Do not know= 36799 : 1299 : 124 : 547 because we also add another label (*False*) for all the *Non-relevant* sentences. This poses a challenge for the JMs as this unbalancing is quite significant. In order to tackle this problem, we employ two techniques.

4.5.1 Weighted Loss for Class Imbalance. We implemented weighted loss technique [29] where the total loss is calculated as weighted sum of loss according to the class. Log weighting helps in smoothing the weights for highly unbalanced classes, which is the case in our dataset as shown in Table 2

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

Where $q \in \{2, 3, 5, 7, 10\}$ represents one of the 5 questions and $c \in$ {Yes, No, Do not know, False} and we tuned α as a hyperparamter which came out to be = 15. T_q is the count for question q and $T_{c,q}$ is the count of a particular class c for question q.

4.5.2 Down-sampling. In order to level the training data, one of the most common technique is to down-sample the datapoints of the most frequent labels. In our case, *False* label has incredibly large amount of data when compared to other three labels. Hence, we employed down-sampling on data for *False* label during training. Before each training epoch, we randomly reduced the training data for *False* label by X. We tuned X as a hyperparameter for the model, which came out to be $X = 40\%$.

For experiments, we applied *weighted loss* technique to all the models, but we only applied *down-sampling* to the *Joint Model with doc representation* since we aimed to observe the incremental value

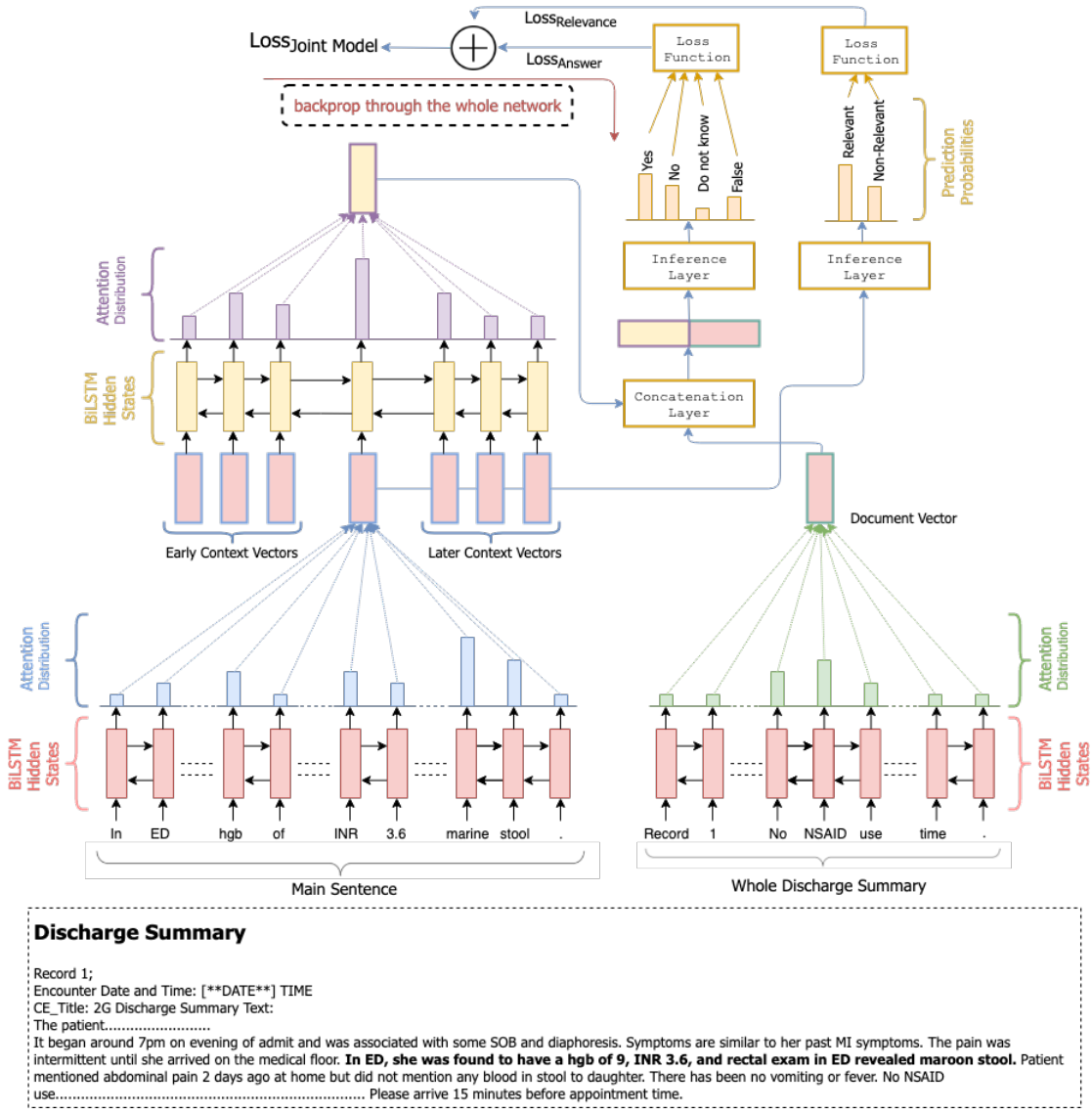

Figure 1: Joint Model with document representation. The first attention layer attends to the words in a sentence. The second attention layer attends to the sentence and it's neighbouring sentences which act as the *local context* for the sentence. The document vector is created separately which provides *global* context to the model. The BiLSTM parameters shared between different parts of the model are color-coded so that the same color represents shared parameters

of *down-sampling*. This model has been referred to as *JM-Doc down* model in the further sections.

4.6 Evaluation Metrics

We evaluated our models on precision, recall and f-score metrics as the answer for each question is limited to four classes: Yes, No, Do not know, False. We reported both macro-averaged and micro-averaged precision recall and f-score for all models 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 *False* label. Macro-averaged metrics are calculated by averaging the performance across the labels and thus provide better insight on model's performance across different labels.

5 RESULTS AND ANALYSIS

We reported both macro-averaged and micro-averaged precision, recall and f-score. Micro-average reports the average of instance-level performance and therefore is biased to the label with the larger ratio, while macro-average reports the average performance of different labels.

Table 3: Precision, Recall and F-score for selected Naranjo questions by different models: Pipeline Model, Joint Model (JM), Joint Model with Doc representation (JM-Doc) and Joint Model with Doc representation along with active down-sampling (JM-Doc down). MA and MI denote macro and micro respectively.

Ques #	Model	MA-P	MA-R	MA-F	MI-P	MI-R	MI-F
Ques 2	**Pipeline**	0.3045	0.3185	0.3105	0.9313	0.9313	0.9313
	JM	0.4445	0.4472	0.4424	0.9545	0.9545	0.9545
	JM-Doc	0.4608	0.4677	0.4633	0.9592	0.9592	0.9592
	JM-Doc down	0.4874	0.4952	**0.4827**	0.9624	0.9624	**0.9624**
Ques 3	**Pipeline**	0.3657	0.3776	**0.3675**	0.9809	0.9809	0.9809
	JM	0.3459	0.3423	0.3434	0.9902	0.9902	0.9902
	JM-Doc	0.6640	0.3381	0.3415	0.9918	0.9918	**0.9918**
	JM-Doc down	0.3546	0.4007	0.3652	0.9780	0.9780	0.9780
Ques 5	**Pipeline**	0.3137	0.3302	0.3209	0.9313	0.9313	0.9313
	JM	0.3791	0.4181	0.3884	0.9404	0.9404	0.9404
	JM-Doc	0.3722	0.3907	0.3758	0.9434	0.9434	0.9434
	JM-Doc down	0.4054	0.3859	**0.3936**	0.9523	0.9523	**0.9523**
Ques 7	**Pipeline**	0.2785	0.3070	0.2876	0.9728	0.9728	0.9728
	JM	0.2890	0.3523	0.3054	0.9694	0.9694	0.9694
	JM-Doc	0.3838	0.3558	**0.3678**	0.9874	0.9874	**0.9874**
	JM-Doc down	0.3587	0.3585	0.3409	0.9858	0.9858	0.9858
Ques 10	**Pipeline**	0.3275	0.3274	0.3260	0.9288	0.9288	0.9288
	JM	0.5017	0.4826	0.4886	0.9535	0.9535	0.9535
	JM-Doc	0.5104	0.4628	0.4779	0.9542	0.9542	**0.9542**
	JM-Doc down	0.5394	0.5365	**0.5271**	0.9494	0.9494	0.9494

Based on micro-averaged f-scores, the JM outperformed the pipeline model on all five questions. In contrast, based on the macro-averaged f-score, the joint models perform better than the pipeline model in four out of the five selected questions. In question 3, the Pipeline model (macro F1=0.3675) performs slightly better than the joint models (macro F1=0.3652), but the JM outperformed the Pipeline model in micro F1 scores (0.9918 vs 0.9809). As shown in Table 3, the joint models with document representation generally outperform the models that only uses local information, i.e., sentence and its context. We provide further analyses for each question, as follow.

5.1 Question 2

Did the adverse event occur after the suspected drug was administered?

All three joint models perform better than the pipeline model based on macro-averaged as well as micro-averaged metrics. The joint model with the doc representations performs better than other joint models. Our hypothesis is that document representation helps the model in learning difficult patterns of the sentences as well. We analyzed multiple sentences to understand the patterns where the JM models with document representation perform better. An example for a sentence is shown below:

> *Sentence:* Never had bleeding like this before.
> *Answer:* Yes.

In the sentence above, the patient experiences *bleeding* after the anticoagulant is administered to the patient hence the answer for question 2 is 'Yes'.

The variant of joint models with document representation predicts the correct answer for this sentence whereas the two other models got it wrong. In this EHR note, the information that the anticoagulant is administered to the patient is mentioned quite early in the discharge summary and is not captured in the context sentences of the above sentence. The document representation helps in summarizing this information. This is also validated in Fig. 2 which shows how the token level attention varies over different words. The pipeline model and joint model focuses attention over different word than *bleeding*. The joint model most likely got confused because of the presence of the word *never* at the start of the sentence because of its' negative connotation. Whereas the two variants of joint models with document representation focuses its attention on the *bleeding* event which is an ADR for the anticoagulant. We observed such similar patterns of attention shift in other examples as well.

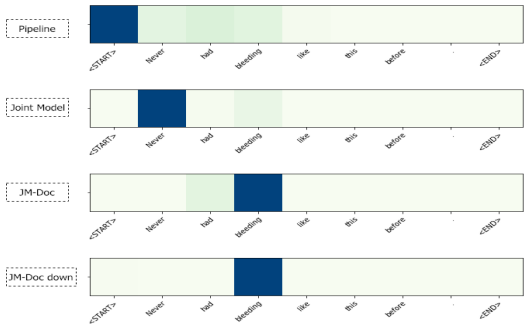

Figure 2: Token level attention of a sentence for question 2.

5.2 Question 3

Did the adverse reaction improve when the drug was discontinued or a specific antagonist was administered?

For this question, the performance of the pipeline model is the best when we look at the macro-averaged classification metrics, although the difference in the macro-weighted f-score is small (0.026) and not statistically significant. In contrast, the JM has the highest performance based on micro-weighted f-score (0.9918), which is more than 0.01 higher than the pipeline model.

The mix results may be caused by the fact that question 3 has the least amount of data for training (281, 21, 181 for Yes, No, and Do not know, respectively), which is in contrast to other questions (for example, question 10 has 1683, 678, and 227 for Yes, No, and Do not know). Among all five questions, question 3 has the least frequency counts because most of the time the patient's visit is not long enough to have a conclusive answer regarding whether the ADR improved once the drug was discontinued. It is usually a suggestion to patients during discharge to discontinue the medicine or use an antagonist along with it. Hence, it is unlikely in the discharge summary to suggest whether the patient observed any improvement or not.

5.3 Question 5

Are there alternative causes that could have on other own cause the reaction?

This question has the highest number for training (2186, 221, and 316 for Yes, No and Do not know) and yet the performance (the highest macro-F1 score is 0.3936) is not highest compared to other Naranjo questions. The joint model with document representation and down-sampling performs the best for both macro and micro-weighted f-score. After manually examining multiple examples, we observed certain patterns (e.g., triggers that can cause the ADRs) can be mentioned in negative connotation as well as positive connotation in the discharge summary and the models may not have sufficient amount of data, even though the training size of this question is the highest among all the Naranjo questionnaire. Two examples are shown below to illustrate this problem in more detail.

> *Statement:* No history of diverticulosis , has never had a colonoscopy.
> *Answer:* No.

Diverticulosis is the condition of having multiple pouches in the colon that are not generally inflammed. This can cause painless rectal and is usually the main cause of lower gastrointestinal bleeding. Since the patient does not have a history of diverticular disease which rules out the possibility that it could have acted like a trigger or alternative cause for ADR, hence the answer is marked as 'No'.

None of the models were able to predict the correct answer for this sentence. In Fig. 3, all the models focused their attention over tokens such as *diverticulosis* and *colonoscopy* which can confuse the model as trigger for the ADR, if the negative connotation for that trigger is not considered with the help of the first token of the sentence: *No.*

> *Statement:* Impression: - Diverticulosis sigmoid colon, descending colon and ascending colon.
> *Answer:* Yes.

In the above example, the physician observes that there are diverticulosis colons in the patient. As mentioned earlier, diverticulosis colons can result in rectal bleeding which is why the answer for this sentence with respect to question 5 is Yes.

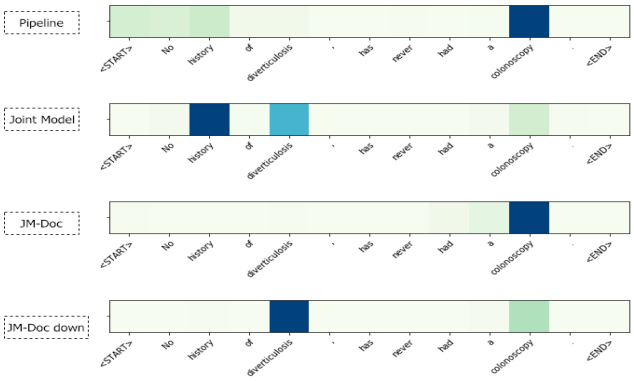

Figure 3: Token level attention of a sentence for question 5.

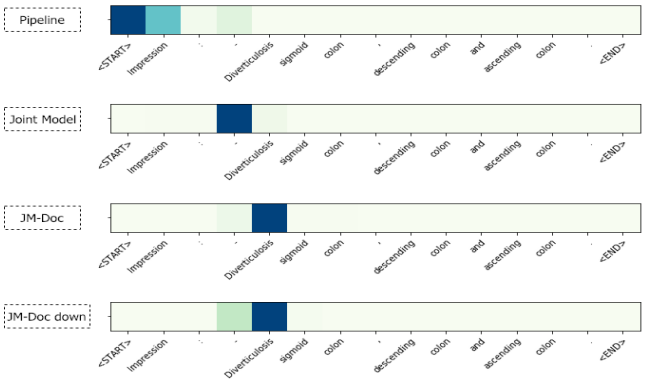

Figure 4: Token level attention of a sentence for question 5.

All the joint models were able to correctly predict the answer for this sentence with respect to question 5. We can observe in Fig. 4 that except the Pipeline model all the other models were able to focus their attention on the token *diverticulosis* and predicting the correct answer accordingly after recognizing the trigger for bleeding.

Here we saw two examples for question 5 which had the same token, *diverticulosis*, but according to the context the answers were different. The models learned to identify the triggers in the sentence but were not able to learn the negative connotation mentioned along with it. This problem can easily occur because of less data.

5.4 Question 7

Was the drug detected in the blood (or other fluids) in concentration known to be toxic?

Similar to other questions, this question also presents a data imbalance challenge. In Table 3, all three joint models outperformed the pipeline model. Among the JMs, JM-Doc performed the best. We analyzed different datapoints of wrongly predicted answers, and found that the models are not able to capture compound information. Examples are shown below:

Statement: Pt required multiple blood transfusions daily and had many transfusions of FFP to achieve a hgb > 8 and INR < 1.5.
Answer: Yes.

The above statement suggests that the patient was given multiple transfusions to get his/her INR (international normalized ratio) below 1.5. INR test values are used to determine the clotting tendencies of the blood and the normal values are usually below 1.5.

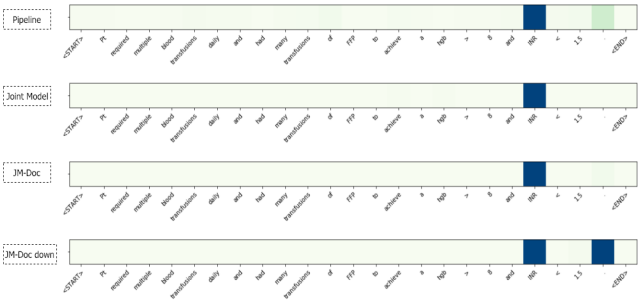

Figure 5: Token level attention of a sentence for question 7.

All of the models predicted the answer for this statement with respect to question 7 correctly. The above figure shows that all the models were able to focus their attention on the INR token which is provided in the statement and predicted the 'Yes' token accordingly. Though it should be noted that none of the models fixed their attention on the < sign or the value of INR test (1.5).

Statement: INR was 1.4 , and he was given 80 mg enoxaparin.
Answer: No.

In the statement above, the INR level of the patient are below 1.5 and hence it is within the normal range which is why the answer for this statement is No.

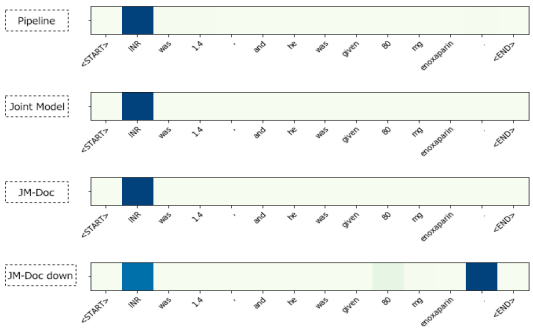

Figure 6: Token level attention of a sentence for question 7.

All the models predicted wrong asnwer for this sentence. When we look at the token level attention in Fig. 6, we observe that all the models focused their attention on the INR token but since the models are not compounding the information of the whole sentence, they end up predicting wrong answer for such sentences.

5.5 Question 10

Was the adverse event confirmed by any objective evidence?

For this question, all the variants of joint models achieved their personal best performance across questions. The joint model with document representation along with downsampling (*JM-Doc down*) performs the best with the macro-weighted f-score of 0.5271. For this question, we tried to analyze where the models failed to predict the correct answers.

Statement: The following day he had a significant increase in his hematuria, and the catheter stopped draining.
Answer: Yes.

Hematuria is the presence of red blood cells in the urine. The increase in hematuria acts as an evidence for the existence of an ADR such as bleeding, hence the answer for this sentence with respect to question 10 is 'Yes'.

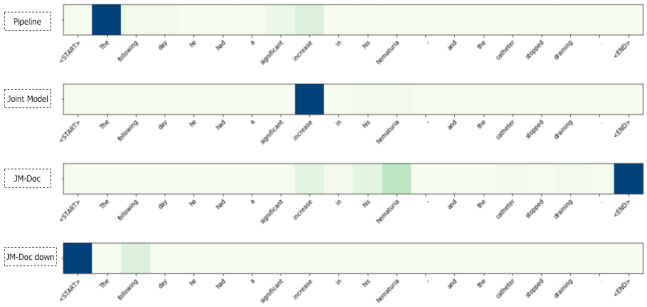

Figure 7: Token level attention of a sentence for question 10.

All the models predicted wrong answer for the sentence mentioned above. In Fig. 7, the attention of all the models is focused on different tokens except *hematuria* which should be the evidence for the adverse event (bleeding). By looking at more examples, it became clear that the models were not able to focus on tokens which have few count in the dataset. Different tokens which do not appear frequently in our dataset tend to be harder to recognize by the model resulting in predicting wrong answers for the sentence. Increasing the data over more patients and discharge summaries would help in boosting the accuracy of models for this question.

Table 4: Model pairs and their t-test p-values.

Pairs	p-value
Pipeline - JM	0.0797
Pipeline - JM-Doc	0.0259*
Pipeline - JM-Doc down	0.0306*

* p<0.05

5.6 Statistical Analysis

In order to check if our results are significant or not, we calculated significance p-values by performing paired t-test on multiple pairs of Pipeline and joint models using their f-score values. For each model pair, we performed the t-test across the questions. As we

can observe in Table 4, the performance improvement of *JM-Doc* and *JM-Doc down* is significant over the performance of Pipeline model as their respective p-values are less than 0.05. The p-value of *Pipeline-JM* pair is quite close to 0.05 and hence the performance improvement of JM over Pipeline can also be considered significant.

6 CONCLUSION

In this paper, we built question answering models for automatically answering the Naranjo questionnaire [22] using EHR notes. The Naranjo scale is well accepted in the medical domain for assessing the causality between a medication and its ADRs. We built an innovative end-to-end multi-task joint model which generally outperformed the pipeline model for automatically answering the Naranjo questionnaire. We employed different techniques to account for data imbalancing challenge. We also explored different contextual information. Our results show that both the global context (document-level context) and down sampling help improve the performance of joint models. Our study presents the first computation model in EHR-based clinical question answering and in automated Naranjo scale assessment, and therefore would be the baseline model for any future work in these areas.

ACKNOWLEDGMENTS

We would like to thank Drs. Steve Belknap, Feifan Liu, William Temps, and Edgard Granillo and Ms. Nadya Frid for annotating the discharge summaries regarding Naranjo Questionnaire. This work was supported by the grant HL125089 from the National Institutes of Health (NIH). The contents of this paper do not represent the views of NIH.

REFERENCES

[1] R Arulmani, SD Rajendran, and B Suresh. 2008. Adverse drug reaction monitoring in a secondary care hospital in South India. *British journal of clinical pharmacology* 65, 2 (2008), 210–216.
[2] 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. 1997. The costs of adverse drug events in hospitalized patients. *Jama* 277, 4 (1997), 307–311.
[3] Antoine Bordes, Sumit Chopra, and Jason Weston. 2014. Question answering with subgraph embeddings. *arXiv preprint arXiv:1406.3676* (2014).
[4] Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075* (2015).
[5] Rich Caruana. 1997. Multitask learning. *Machine learning* 28, 1 (1997), 41–75.
[6] David C Classen, Stanley L Pestotnik, R Scott Evans, James F Lloyd, and John P Burke. 1997. Adverse drug events in hospitalized patients: excess length of stay, extra costs, and attributable mortality. *Jama* 277, 4 (1997), 301–306.
[7] David C Classen, Roger Resar, Frances Griffin, Frank Federico, Terri Frankel, Nancy Kimmel, John C Whittington, Allan Frankel, Andrew Seger, and Brent C James. 2011. âĂŸGlobal trigger toolâĂŹshows that adverse events in hospitals may be ten times greater than previously measured. *Health affairs* 30, 4 (2011), 581–589.
[8] EC Davies, CF Green, DR Mottram, and M Pirmohamed. 2006. Adverse drug reactions in hospital in-patients: a pilot study. *Journal of clinical pharmacy and therapeutics* 31, 4 (2006), 335–341.
[9] Emma C Davies, Christopher F Green, Stephen Taylor, Paula R Williamson, David R Mottram, and Munir Pirmohamed. 2009. Adverse drug reactions in hospital in-patients: a prospective analysis of 3695 patient-episodes. *PLoS one* 4, 2 (2009), e4439.
[10] Ronen Feldman, Oded Netzer, Aviv Peretz, and Binyamin Rosenfeld. 2015. 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*. ACM, 1779–1788.
[11] Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18, 5-6 (2005), 602–610.
[12] Rave Harpaz, Krystl Haerian, Herbert S Chase, and Carol Friedman. 2010. Mining electronic health records for adverse drug effects using regression based methods. In *Proceedings of the 1st ACM International Health Informatics Symposium*. ACM, 100–107.
[13] Aron Henriksson, Maria Kvist, Hercules Dalianis, and Martin Duneld. 2015. Identifying adverse drug event information in clinical notes with distributional semantic representations of context. *Journal of biomedical informatics* 57 (2015), 333–349.
[14] Trung Huynh, Yulan He, Alistair Willis, and Stefan Rüger. 2016. Adverse drug reaction classification with deep neural networks. Coling.
[15] Abhyuday N Jagannatha and Hong Yu. 2016. Bidirectional rnn for medical event detection in electronic health records. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, Vol. 2016. NIH Public Access, 473.
[16] Daniel R Levinson and Inspector General. 2010. Adverse events in hospitals: national incidence among Medicare beneficiaries. *Department of Health and Human Services Office of the Inspector General* (2010).
[17] Fei Li, Weisong Liu, and Hong Yu. 2018. Extraction of information related to adverse drug events from electronic health record notes: design of an end-to-end model based on deep learning. *JMIR medical informatics* 6, 4 (2018), e12159.
[18] Jennifer Lucado, Kathryn Paez, and A Elixhauser. 2006. Medication-related adverse outcomes in US hospitals and emergency departments, 2008: statistical brief# 109. (2006).
[19] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).
[20] Sewon Min, Minjoon Seo, and Hannaneh Hajishirzi. 2017. Question answering through transfer learning from large fine-grained supervision data. *arXiv preprint arXiv:1702.02171* (2017).
[21] Tsendsuren Munkhdalai, Feifan Liu, and Hong Yu. 2018. 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).
[22] Cláudio A Naranjo, Usoa Busto, Edward M Sellers, P Sandor, I Ruiz, EA Roberts, E Janecek, C Domecq, and DJ Greenblatt. 1981. A method for estimating the probability of adverse drug reactions. *Clinical Pharmacology & Therapeutics* 30, 2 (1981), 239–245.
[23] Maria Cristina G Passarelli, Wilson Jacob-Filho, and Albert Figueras. 2005. Adverse drug reactions in an elderly hospitalised population. *Drugs & aging* 22, 9 (2005), 767–777.
[24] Walter W Piegorsch. 1990. Maximum likelihood estimation for the negative binomial dispersion parameter. *Biometrics* (1990), 863–867.
[25] R Priyadharsini, A Surendiran, C Adithan, S Sreenivasan, and Firoj Kumar Sahoo. 2011. A study of adverse drug reactions in pediatric patients. *Journal of pharmacology & pharmacotherapeutics* 2, 4 (2011), 277.
[26] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603* (2016).
[27] Parikshit Sondhi, Jimeng Sun, Hanghang Tong, and ChengXiang Zhai. 2012. 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*. ACM, 1167–1175.
[28] Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. 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)*, Vol. 1. 189–198.
[29] Zhi-Hua Zhou and Xu-Ying Liu. 2006. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering* 18, 1 (2006), 63–77.