

Entity-Enriched Neural Models for Clinical Question Answering

Bhanu Pratap Singh Rawat^{1,*}, Wei-Hung Weng^{2,3,#}, So Yeon Min^{2,3,§},

Preethi Raghavan^{3,4,†}, Peter Szolovits^{2,3,‡}

¹UMass-Amherst, ²MIT CSAIL, ³MIT-IBM Watson AI Lab, ⁴IBM Research, Cambridge

*brawat@umass.edu, #ckbjimmy@mit.edu, §symin95@mit.edu

†praghav@us.ibm.com, ‡psz@mit.edu

Abstract

We explore state-of-the-art neural models for question answering on electronic medical records and improve their ability to generalize better on previously unseen (paraphrased) questions at test time. We enable this by learning to predict logical forms as an auxiliary task along with the main task of answer span detection. The predicted logical forms also serve as a rationale for the answer. Further, we also incorporate medical entity information in these models via the ERNIE (Zhang et al., 2019a) architecture. We train our models on the large-scale emrQA dataset and observe that our multi-task entity-enriched models generalize to paraphrased questions ~ 5% better than the baseline BERT model.

1 Introduction

The field of question answering (QA) has seen significant progress with several resources, models and benchmark datasets. Pre-trained neural language encoders like BERT (Devlin et al., 2019) and its variants (Seo et al., 2016; Zhang et al., 2019b) have achieved near-human or even better performance on popular open-domain QA tasks such as SQuAD 2.0 (Rajpurkar et al., 2016). While there has been some progress in biomedical QA on medical literature (Šuster and Daelemans, 2018; Tsatsaronis et al., 2012), existing models have not been similarly adapted to clinical domain on electronic medical records (EMRs).

Community-shared large-scale datasets like emrQA (Pampari et al., 2018) allow us to apply state-of-the-art models, establish benchmarks, innovate and adapt them to clinical domain-specific needs. emrQA enables question answering from electronic medical records (EMRs) where a question is asked by a physician against a patient’s medical record

Context: The patient had an elective termination of her pregnancy on [DATE]. The work-up for the extent of the patient’s disease included mri scan of the cervical and thoracic spine which revealed multiple metastatic lesions in the vertebral bodies; A T 3 lesion extending from the body to the right neural foramina with foraminal obstruction. **An abdominal and pelvic ct scan with iv contrast** revealed bilateral pulmonary nodules and bilateral pleural effusions, extensive liver metastases, narrowing of the intra hepatic ivc and distention of the azygous system suggestive of ivc obstruction by liver metastases.

Question: How was the patient’s extensive liver metastases diagnosed?

Paraphrase: What diagnosis was used for the patient’s extensive liver metastases?

Logical Form: {LabEvent (x) [date=x, result=x] OR ProcedureEvent (x) [date=x, result=x] OR VitalEvent (x) [date=x, result=x]} reveals ConditionEvent ([problem])

Answer: An abdominal and pelvic ct scan with iv contrast

Figure 1: A synthetic example of a clinical context, question, its logical form and the expected answer.

(clinical notes). Thus, we adapt these models for EMR QA while focusing on model generalization via the following. (1) learning to predict the logical form (a structured semantic representation that captures the answering needs corresponding to a natural language question) along with the answer and (2) incorporating medical entity embeddings into models for EMR QA. We now examine the motivation behind these.

A physician interacting with a QA system on EMRs may ask the same question in several different ways; a physician may frame a question as: “Is the patient allergic to penicillin?” whereas the other could frame it as “Does penicillin cause any allergic reactions to the patient?”. Since paraphrasing is a common form of generalization in natural language processing (NLP) (Bhagat et al., 2009), a QA model should be able to generalize well to such paraphrased question variants that may not be seen during training (and avoid simply memorizing the questions). However, current state-of-the-art models do not consider the use of meta-information such as the semantic parse or logical form of the questions in unstructured QA. In order to give the model the ability to understand the semantic information about answering needs of a question, we frame our problem in a multitask learning setting where the primary task is extractive QA and the

*The author did this work while interning at MIT-IBM Watson AI Lab.

auxiliary task is the logical form prediction of the question.

Fine-tuning on medical corpora (MIMIC-III, PubMed (Johnson et al., 2016; Lee et al., 2020)) helps models like BERT align their representations according to medical vocabulary (since they are previously trained on open-domain corpora such as WikiText (Zhu et al., 2015)). However, another challenge for developing EMR QA models is that different physicians can use different medical terminology to express the same entity; e.g., “heart attack” vs. “myocardial infarction”. Mapping these phrases to the same UMLS semantic type¹ as *disease or syndrome* (*dsyn*) provides common information between such medical terminologies. Incorporating such entity information about tokens in the context and question can further improve the performance of QA models for the clinical domain.

Our contributions are as follows:

1. We establish state-of-the-art benchmarks for EMR QA on a large clinical question answering dataset, emrQA (Pampari et al., 2018)
2. We demonstrate that incorporating an auxiliary task of predicting the logical form of a question helps the proposed models generalize well over unseen paraphrases, improving the overall performance on emrQA by $\sim 5\%$ over BERT (Devlin et al., 2019) and by $\sim 3.5\%$ over clinicalBERT (Alsentzer et al., 2019). We support this hypothesis by running our proposed model over both emrQA and another clinical QA dataset, MADE (Jagannatha et al., 2019).
3. The predicted logical form for unseen paraphrases helps in understanding the model better and provides a rationale (explanation) for why the answer was predicted for the provided question. This information is critical in *clinical domain* as it provides an accompanying answer justification for clinicians.
4. We incorporate medical entity information by including entity embeddings via the ERNIE (Zhang et al., 2019a) architecture (Zhang et al., 2019a) and observe that the model accuracy and ability to generalize goes up by $\sim 3\%$ over BERT_{base} (Devlin et al., 2019).

¹<https://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml>

2 Problem Formulation

We formulate the EMR QA problem as a reading comprehension task. Given a natural language *question* (asked by a physician) and a *context*, where the context is a set of contiguous sentences from a patient’s EMR (unstructured clinical notes), the task is to predict the answer span from the given context. Along with the (question, context, answer) triplet, also available as input are *clinical entities* extracted from the question and context. Also available as input is the, *logical form* (LF) that is a structured representation that captures answering needs of the question through entities, attributes and relations required to be in the answer (Pampari et al., 2018). A question may have multiple paraphrases where all paraphrases map to the same LF (and the same answer, fig. 1).

3 Methodology

In this section, we briefly describe BERT (Devlin et al., 2019), ERNIE (Zhang et al., 2019a) and our proposed model.

3.1 Bidirectional Encoder Representations from Transformers (BERT)

BERT (Devlin et al., 2019) uses multi-layer bidirectional Transformer (Vaswani et al., 2017) networks to encode contextualised language representations. BERT representations are learned from two tasks: masked language modeling (Taylor, 1953) and next sentence prediction task. We chose BERT model as pre-trained BERT models can be fine-tuned with just one additional inference layer and it achieved state-of-the-art results for a wide range of tasks such as question answering, such as SQuAD (Rajpurkar et al., 2016, 2018), and multiple language inference tasks, such as MultiNLI (Williams et al., 2017). *clinicalBERT* (Alsentzer et al., 2019) yielded superior performance on clinical-related NLP tasks such as i2b2 named entity recognition (NER) challenges (Uzuner et al., 2011). It was created by further fine-tuning of BERT_{base} with biomedical and clinical corpus (MIMIC-III) (Johnson et al., 2016).

3.2 Enhanced Language Representation with Informative Entities (ERNIE)

We adopt the ERNIE framework (Zhang et al., 2019a) to integrate the entity-level clinical concept information into the BERT architecture, which has not yet been explored in the previous works.

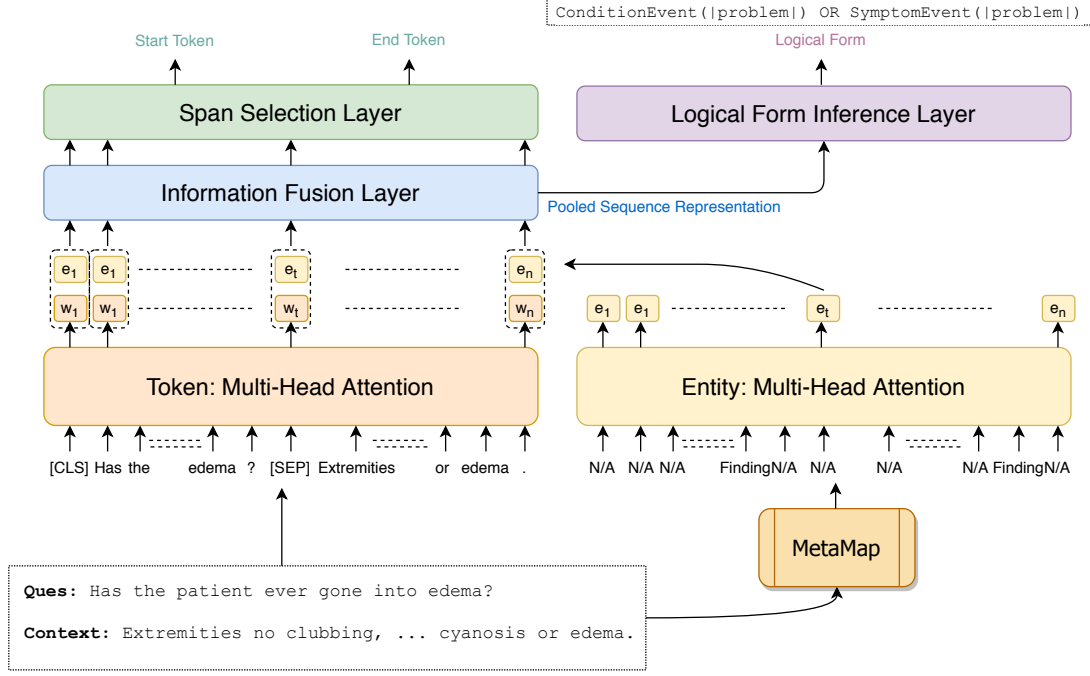


Figure 2: The network architecture of our multi-task learning question answering model (*M-cERNIE*). The question and context are provided to a multi-head attention model (orange) and are also passed through MetaMap to extract clinical entities which are passed through a separate multi-head attention (yellow). The token and entity representations are then passed through an information fusion layer (blue) to extract entity-enriched token representations which are then used for answer span prediction. The pooled sequence representation from the information fusion layer is passed through logical form inference layer to predict the logical form.

ERNIE has shown significant improvement in different entity typing and relation classification tasks, as it utilises the extra entity information which is provided from knowledge graphs. ERNIE uses BERT for extracting contextualized token embeddings and a multi-head attention model to generate entity embeddings. These two set of embeddings are aligned and provided as an input to an information fusion layer which provides entity-enriched token embeddings. For a token (w_j) and its aligned entity ($e_k = f(w_j)$), the information fusion process is as follows:

$$h_j = \sigma(W_t^{(i)} w_j^{(i)} + W_e^{(i)} e_k^{(i)} + b^{(i)}) \quad (1)$$

Here h_j represents the entity enriched token embedding, σ is the non-linear activation function, W_t refers to an affine layer for token embeddings and W_e refers to an affine layer for entity embeddings. For the tokens without corresponding entities, the information fusion process becomes:

$$h_j = \sigma(W_t^{(i)} w_j^{(i)} + b^{(i)}) \quad (2)$$

Initially, each entity embedding is assigned randomly and is fine-tuned along with token embeddings throughout the training procedure. The

ERNIE architecture would be applicable to the model even if the logical forms are not available.

3.3 Multi-task Learning for Extractive QA

In order to improve the ability of a QA model to generalize better over paraphrases, it helps to provide the model information about the logical form that links these paraphrases. Since the answer to all the paraphrased questions is the same (and hence, logical form is the same), we constructed a multi-task learning framework to incorporate the logical form information into the model. Thus, along with predicting the answer span, we added an auxiliary task to also predict the corresponding logical form of the question. Multi-task learning provides an inductive bias to enhance the primary task's performance via auxiliary tasks (Weng et al., 2019). In our setting, the primary task is span detection of the answer and the auxiliary task is logical form prediction for both emrQA and MADE (both datasets are explained in detail in § 4). The final loss for our model is defined as:

$$\mathcal{L}_{model} = \omega \mathcal{L}_{lf} + (1 - \omega) \mathcal{L}_{span}, \quad (3)$$

where ω is the weightage given to the loss of auxiliary task (\mathcal{L}_{lf}), logical form prediction. \mathcal{L}_{span}

is loss for answer span prediction and \mathcal{L}_{model} is the final loss for our proposed model. The multi-task learning model can work with both BERT and ERNIE as the base model. Figure 2 depicts the proposed multi-task model to predict both the answer and logical form given a question and ERNIE architecture that is used to learn entity-enriched token embeddings.

4 Datasets

We used *emrQA*² and *MADE*³ datasets for our experiments. We provide a brief summary of each dataset and the methodology followed to split these datasets into train and test sets.

emrQA The *emrQA* corpus (Pampari et al., 2018) is the only community-shared clinical QA dataset that consists of questions, posed by physicians against electronic medical records (EMRs) of a patient, along with their answers. The dataset was developed by leveraging existing annotations available for other clinical natural language processing (NLP) tasks (i2b2 challenge datasets (Uzuner et al., 2011)). It is a credible resource for clinical QA as logical forms that are generated by a physician help slot fill question templates and extract corresponding answers from annotated notes. Multiple question templates can be mapped to the same logical form (LF), as shown in Table 1, and are referred to as *paraphrases* of each other.

LF: MedicationEvent (|medication|) [dosage=x]

How much |medication| does the patient take per day?
 What is her current dose of |medication|?
 What is the current dose of the patient’s |medication|?
 What is the current dose of |medication|?
 What is the dosage of |medication|?
 What was the dosage prescribed of |medication|?

Table 1: A logical form (LF) and its respective question templates (paraphrases).

The *emrQA* corpus has over 1M+ question, logical form, and answer/evidence triplets, an example of a context, question, its logical form and a paraphrase is shown in Fig 1. The evidences are the sentences from the clinical note that are relevant to a particular question. There are total 30 logical forms in the *emrQA* dataset⁴.

²<https://github.com/panushri25/emrQA>

³<https://bio-nlp.org/index.php/projects/39-nlp-challenges>

⁴<https://github.com/panushri25/emrQA/blob/master/templates/templates-all.csv>

MADE *MADE* 1.0 (Jagannatha et al., 2019) dataset was hosted as an adverse drug reactions (ADRs) and medication extraction challenge from EMRs. This dataset was converted into a QA dataset by following the same procedure as enumerated in the literature of *emrQA* (Pampari et al., 2018). *MADE* QA dataset is smaller than *emrQA*, as *emrQA* consists of multiple datasets taken from i2b2 (Uzuner et al., 2011) whereas *MADE* only has specific relations and entity mentions to that of ADRs and medications. This resulted in a clinical QA dataset which has different properties as compared to *emrQA*. *MADE* also has lesser number of logical forms (8 LFs) as compared to *emrQA* because of fewer entities and relations. The 8 LFs for *MADE* are provided in Appendix B.

4.1 Train/test splits

The *emrQA* dataset is generated using a semi-automated process that normalizes real physician questions to create question templates, associates expert annotated logical forms with each template and slot fills them using annotations for various NLP tasks from i2b2 challenge datasets (for e.g., fig. 1). *emrQA* is rich in paraphrases as physicians often tend to express the same information need in different ways. As shown in Table. 1, all paraphrases of a question map to the same logical form. Thus, if a model has observed some of the paraphrases it should be able to generalize to the others effectively with the help of their shared logical form “MedicationEvent (|medication|) [dosage=x]”. In order to simulate this, and test the true capability of the model to generalize to unseen paraphrased questions, we create a splitting scheme and refer to it as *paraphrase-level split*.

Paraphrase-level split

The basic idea is that some of question templates would be observed by the model during training and remaining would be used during validation and testing. The steps taken for creating this split are enumerated below:

1. First, the clinical notes are separated into train, val and test sets. Then the question, logical form and context triplets are generated for each set resulting in the full dataset. Here the context is the set of contiguous sentences from the EMR.
2. Then for each logical form (LF), 70% of its

corresponding question templates are chosen for train dataset and the rest are kept for validation and test dataset. Considering the LF shown in Table 1, four of the question templates (QT_{tr}) would be assigned for training and two ($QT_{v/t}$) of them would be assigned for validation/testing. So any sample in training dataset whose question is generated from the question template set $Q_{v/t}$ would be discarded. Similarly, any sample with a question generated from the question template set Q_{tr} would be discarded.

3. To compare the generalizability performance of our model, we keep the training dataset with both set of question templates ($QT_{tr} + QT_{v/t}$) as well. Essentially, a baseline model which has observed all the question templates ($QT_{tr} + QT_{v/t}$) should be able to perform better on the $QT_{v/t}$ set as compared to a model which has only observed QT_{tr} set. This comparison would help us in measuring the improvement in performance with the help of logical forms even when a set of question templates are not observed by the model.

The dataset statistics for both *emrQA* and *MADE* are shown in Table 2. The training set with both question template sets ($QT_{tr} + QT_{v/t}$) is shown with ‘(r)’ appended as suffix, as it is essentially a *random* split, whereas the training set with the question template (QT_{tr}) is appended with ‘(pl)’ for *paraphrase-level* split.

Datasets	Split	Train	Val.	Test
emrQA	# Notes	433	44	47
	# Samples (pl)	133,589	21,666	19,401
	# Samples (r)	198,118	21,666	19,401
MADE	# Notes	788	88	213
	# Samples (pl)	73,224	4,806	9,235
	# Samples (r)	113,975	4,806	9,235

Table 2: Train, validation and test data splits.

5 Experiments

In this section, we briefly discuss the experimental settings, clinical entity extraction method, implementation details of our proposed model and evaluation metrics for our experiments.

5.1 Experimental Setting

As a reading comprehension style task, the model has to identify the span of the answer given the

question-context pair. For both *emrQA* and *MADE* dataset, the span is marked as the answer to the question and the sentence is marked as the evidence. Hence, we perform extractive question answering at two levels: *sentence* and *paragraph*.

Sentence setting: For this setting, the evidence sentence which contains the answer span is provided as the context to the question and the model has to predict the span of the answer, given the question.

Paragraph setting: Clinical notes are noisy and often contain incomplete sentences, lists and embedded tables making it difficult to segment paragraphs in notes. Hence, we decided to define the context as evidence sentence and 15 – 20 sentences around it. We randomly chose the length of the paragraph (l_{para}) and another number less than the length of the paragraph ($l_{pre} < l_{para}$). We chose l_{pre} contiguous sentences which exist prior to the evidence sentence in the EMR and $(l_{para} - l_{pre})$ sentences after the evidence sentence. We adopted this strategy because the model could have benefited from the information that the evidence sentence is exactly in the middle of a fixed length paragraph. The model has to predict the span of the answer from the l_{para} sentences long paragraph (context) given the question.

The datasets are appended by ‘-p’ and ‘-s’ for *paragraph* and *sentence* settings respectively. The *sentence* setting is a relatively easier setting, for the model, compared to the *paragraph* setting because the scope of the answer is narrowed down to lesser number of tokens and there is less noise. For both settings, as also mentioned in § 4, we kept the train set where all the question templates (paraphrases) are observed by the model during training and that is referred with ‘(r)’ prefix, suggesting ‘random’ selection and no filtering based on question templates (paraphrases). All these dataset abbreviations are shown in the first column of Table 3.

5.2 Extracting Entity Information

MetaMap (Aronson, 2001) uses a knowledge-intensive approach to discover different clinical concepts referred to in the text according to unified medical language system (UMLS) (Bodenreider, 2004). The clinical ontologies, such as SNOMED (Spackman et al., 1997) and RxNorm (Liu et al., 2005), embedded in MetaMap are quite useful in extracting ~ 127 entities across diagnosis, medication, procedure and sign/symptoms. We shortlisted

these entities (semantic types) by mapping them to the entities which were used for creating logical forms of the questions as these are the main entities for which the question has been posed. The selected entities are: acab, aggp, anab, anst, bpoc, cgab, clnd, diap, emod, evnt, fndg, inpo, lbpr, lbtr, phob, qnco, sbst, sosy and topp. Their descriptions are provided in Appendix C.

These filtered entities (Table 7), extracted from MetaMap, are provided to ERNIE. A separate embedding space is defined for the entity embeddings which are passed through a multi-head attention layer (Vaswani et al., 2017) before interacting with token embeddings in the information fusion layer. The entity-enriched token embeddings are then used to predict the span of the answer from the context. We fine-tuned these entity embeddings along with the token embeddings, as opposed to using learned entities and not fine-tuning during downstream tasks (Zhang et al., 2019a). The architecture is illustrated in Fig 2.

5.3 Implementation Details

The BERT model was released with pre-trained weights as BERT_{base} and BERT_{large}. BERT_{base} has lesser number of parameters but achieved state-of-the-art results on a number of open-domain NLP tasks. We performed our experiments with BERT_{base} and hence, from here onwards we refer to BERT_{base} as BERT. A fine-tuned version of BERT_{base} on clinical notes was released as clinicalBERT (cBERT) (Alsentzer et al., 2019). We use cBERT as the multi-head attention model for getting the token representations in ERNIE. We refer to this version of ERNIE, with entities from MetaMap, as cERNIE for clinical ERNIE. Our final multi-task learning model, incorporated with an auxiliary task of predicting logical forms, is referred to as M-cERNIE for multi-task clinical ERNIE. The code for all the models is provided at https://github.com/emrQA/bionlp_acl20.

Evaluation Metrics For our extractive question answering task, we utilised exact match and F1-score for evaluation as per earlier literature (Rajpurkar et al., 2016).

6 Results and Discussion

In this section, we compare the results of all the models that we introduced in § 3. With the help of different experiments, we try to analyse whether the induced entity and logical form information

Dataset	Model	F1-score	Exact Match
emrQA-s (pl)	BERT	72.13	65.81
	cBERT	74.75 (+2.62)	67.25 (+1.44)
	cERNIE	77.39 (+5.26)	70.17 (+4.36)
	M-cERNIE	79.87 (+7.74)	71.86 (+6.05)
emrQA-s (r)	cBERT	82.34	74.58
emrQA-p (pl)	BERT	64.19	56.30
	cBERT	65.45 (+1.26)	57.58 (+1.28)
	cERNIE	66.15 (+1.96)	59.80 (+3.5)
	M-cERNIE	67.21 (+3.02)	61.22 (+4.92)
emrQA-p (r)	cBERT	72.51	65.14
MADE-s (pl)	BERT	68.45	60.73
	cBERT	70.19 (+1.74)	62.00 (+1.27)
	cERNIE	71.51 (+3.06)	65.31 (+4.58)
	M-cERNIE	73.83 (+5.38)	67.53 (+6.8)
MADE-s (r)	cBERT	73.70	65.54
MADE-p (pl)	BERT	63.39	57.49
	cBERT	64.97 (+1.58)	58.94 (+1.45)
	cERNIE	65.71 (+2.32)	60.55 (+3.06)
	M-cERNIE	64.58 (+1.19)	59.39 (+1.9)
MADE-p (r)	cBERT	66.89	61.27

Table 3: F1-score and exact match values for Models on *emrQA* and *MADE*. The ‘-s’ suffix refers to the *sentence* setting and ‘-p’ refers to the *paragraph* setting for the context provided in our reading comprehension style QA task. The ‘(pl)’ refers to the *paraphrase-level* and ‘(r)’ refers to the *random* split as explained in § 4. BERT refers to BERT_{base}, cBERT refers to *clinicalBERT*, cERNIE refers to *clinicalERNIE* and M-cERNIE refers to the multi-task learning *clinicalERNIE* model.

help the model in achieving better performance or not. We also analyse the logical form predictions to understand whether it provides a rationale for the answer predicted by our proposed model. The compiled results for all the models are shown in Table 3. The hyper-parameter values for the best performing models are provided in Appendix A.

Does clinical entity information improve models’ performance?

Across all settings, the F1-score of cERNIE improves by $\sim 2 - 5\%$ over BERT and $\sim 0.75 - 3\%$ over cBERT. The exact match performance improved by $\sim 3 - 4.5$ over BERT and $1.5 - 3.25\%$ over cBERT. Also, as expected, the performance in *sentence setting* (-s) improved relatively more than it did in *paragraph-setting*. The entity-enriched tokens help in identifying the tokens which are required by the question. For example, in Fig. 3, the token ‘infiltrative’ in the question as well as the context get highlighted with the help of the identified entity ‘topp’ (therapeutic or preventive procedure) and then relevant tokens in the context, chest x ray, get highlighted with the relevant entity ‘diap’ (diagnostic procedure). This

information aids the model in narrowing down its focus to highlighted diagnostic procedures in the context for answer extraction.

Question: How was diffuse *infiltrative_{topp}* *diagnosed_{mdg}*?
Context: Earlier that day, pt had a *chest x ray_{diag}* which showed diffuse *infiltrative_{topp}* process concerning for ARDS.
Answer: chest x ray

Figure 3: An example of a question, context, their extracted entities and expected answer.

Does logical form information help the model generalize better? In order to answer this question, we compared the performance of our *M-cERNIE* model to *cERNIE* model and observed an improvement of 1.1 – 2.5% in F1-score and an improvement of 1.4 – 1.8% in exact match performance. Here as well, the performance improvement is more for *sentence setting (-s)* as compared to the *paragraph setting (-p)*. This helps the model in understanding the information need expressed in the question and helps in narrowing down its focus to certain tokens as the candidate answer. As seen in example 3, the logical form helps in understanding that the ‘dose’ of ‘medication’ needs to be extracted from the context where ‘dose’ was already highlighted with the help of the entity embedding of ‘qncq’.

Overall, the performance of our proposed model improves the F1-score by 1.2 – 7.7% and exact-match by 3.1 – 6.8% over *BERT* model. Thus, embedding clinical entity information with the help of further fine-tuning, entity-enriching and logical form prediction help the model in performing better over the unseen paraphrases by a significant margin. For *emrQA*, the performance of *M-cERNIE* is still below the upper bound performance of the *cBERT* model which is achieved when all the question templates are observed (*emrQA-s/p (r)*) by the model but for *MADE*, in *sentence setting (-s)*, the performance of *M-cERNIE* is even better than the upper bound model performance. For *MADE-p* dataset the performance dropped a little when the LF prediction information is added to the model which might be because *MADE-p* only has 8 logical forms (Appendix B) in total, resulting in low variety between the questions. Thus, the auxiliary task did not add much value to the learning of the base model (*cERNIE*) at paragraph level.

Does the model provide a supporting rationale via logical form (LF) prediction? We analyzed

the performance of *M-cERNIE* on *MADE-s* and *emrQA-s* datasets for logical form prediction, as we saw most improvement in *sentence setting (-s)*. We calculated macro-weighted precision, recall and F1-score for logical form classification. The model achieved a F1-score of $\sim 0.45 - 0.59$ for both datasets, as shown in Table 4, *exact match setting*. We analysed the confusion matrix of predicted LF and observed that the model mainly gets confused between the logical forms which convey similar semantic information as shown in Fig. 4.

Q1: What were the results of the abnormal BMI on 2094-12-02?
Logical Form: LabEvent (|test|) [abnormalResultFlag=Y, date=|date|, result=x] OR ProcedureEvent (|test|) [abnormalResultFlag=Y, date=|date|, result=x] OR VitalEvent (|test|) [date=|date|, (result=x)>vital.refhigh] OR VitalEvent (|test|) [date=|date|, (result=x)<vital.reflow] OR [{LabEvent (|test|) [date=|date|, abnormalResultFlag=Y] OR ProcedureEvent (|test|) [date=|date|, abnormalResultFlag=Y] OR VitalEvent (|test|) [date=|date|]} reveals {ConditionEvent (x) OR SymptomEvent (x)}]
Q2: What were the abnormal results of BMI?
Logical Form: LabEvent (|test|) [date=x, (result=x)<lab.reflow] OR LabEvent (|test|) [date=x, (result=x)>lab.refhigh] OR VitalEvent (|test|) [date=x, (result=x)<vital.reflow] OR VitalEvent (|test|) [date=x, (result=x)>vital.refhigh]

Figure 4: Two similar questions with different logical forms (LFs) but overlapping answer conditions.

As we can see in Fig. 4 that both logical forms refer to quite similar information, hence, we decided to obtain performance metrics (precision, recall and F1-score) in *relaxed setting*. We designed this *relaxed setting* to create a more realistic setting, where the *tokens* of predicted and actual logical forms are matched rather than the whole logical form. An example of logical form tokenization is shown in Fig. 5.

LF: MedicationEvent (x) given {ConditionEvent (|problem|) OR SymptomEvent (|problem|)}
Tokenized: ['MedicationEvent (x)', 'given', 'ConditionEvent (|problem|)', 'OR', 'SymptomEvent (|problem|)']

Figure 5: Tokenized logical form (LF).

The model achieves a F1-score of 0.92 for *emrQA-s* and 0.84 for *MADE-s* in relaxed setting (Table 4). This suggests that the model can efficiently identify important semantic information from the question, which is critical for efficient QA. During inference, the *M-cERNIE* models yield a rationale regarding a new test question (unseen paraphrase) by predicting the logical form of the question as an auxiliary task. For ex, the LF in Fig. 1 provides a rationale that any *lab* or *procedure* event related to the *condition* event needs to be extracted from the EMR for diagnosis.

Setting	Dataset	Precision	Recall	F1-score
Exact	emrQA	0.65	0.61	0.59
	MADE	0.47	0.52	0.45
Relaxed	emrQA	0.93	0.91	0.92
	MADE	0.83	0.85	0.84

Table 4: Precision, Recall and F1-score for logical form prediction.

Can logical form information be induced in multi-class QA tasks as well? To answer this question, we performed another experiment where the model has to classify the evidence sentences from the non-evidence sentences making it a two-class classification task. The model would be provided a tuple of question and a sentence and it has to predict whether the sentence is evidence or not? The final loss of the model (\mathcal{L}_{model}) changes to:

$$\mathcal{L}_{model} = \omega \mathcal{L}_{lf} + (1 - \omega) \mathcal{L}_{evidence} \quad (4)$$

where ω is the weightage given to the loss of auxiliary task (\mathcal{L}_{lf}), logical form prediction. $\mathcal{L}_{evidence}$ is loss for evidence classification and \mathcal{L}_{model} is the final loss for our proposed model. We conducted our experiments on *emrQA* dataset as evidence sentences were provided in it. In the multi-class setting, the [CLS] token representation would be used for evidence classification as well as logical form prediction.

Dataset	Model	Precision	Recall	F1-score
emrQA	cBERT	0.67	0.99	0.76
	cERNIE	0.69	0.98	0.78 (+0.02)
	M-cERNIE	0.73	0.99	0.82 (+0.06)

Table 5: Macro-weighted precision, recall and F1-score of Proposed Models on Test Dataset (Multi-choice QA). For the model names, c: clinical; M: multitask.

The multi-task entity enriched model (*M-cERNIE*) achieved an absolute improvement of 6% over *cBERT* and 4% over *cERNIE*. This suggests that the inductive bias introduced via LF prediction does help in improving the overall performance of the model for multi-class QA as well.

7 Related Work

In the general domain, BERT-based models are on the top of different leader boards across various tasks, including QA tasks (Rajpurkar et al., 2018, 2016). The authors of (Nogueira and Cho, 2019) applied BERT to the MS-MARCO passage retrieval QA task and observed improvement over state of the art results. (Nogueira et al., 2019) further extended the work by combining BERT with

re-ranking of predictions for queries that will be issued for each document. However, BERT-based models have not been adapted to answering physician questions on EMRs.

In case of domain-specific QA, logical forms or semantic parse are typically used to integrate the domain knowledge associated with a KB-based (knowledge base) structured QA datasets, where a model is learnt for mapping a natural language question to a LF. GeoQuery (Zelle and Mooney, 1996), and ATIS (Dahl et al., 1994), are the oldest known manually generated question-LF annotations on closed-domain databases. QALD (Lopez et al., 2013), FREE 917 (Cai and Yates, 2013), SIMPLEQuestions (Bordes et al., 2015) contain hundreds of hand-crafted questions and their corresponding database queries. Prior work has also used LFs as a way to generate questions via crowdsourcing (Wang et al., 2015). WEBQuestions (Berant et al., 2013) contains thousands of questions from Google search where the LFs are learned as latent representations in helping answer questions from Freebase. Prior work has not investigated the utility of logical forms in unstructured QA, especially as a means to generalize the QA model across different paraphrases of a question.

There have been efforts on using multi-task learning for efficient question answering, such as the authors of (McCann et al., 2018) tried to learn multiple tasks together resulting in an overall boost in the performance of the model on SQuAD (Rajpurkar et al., 2016). Similarly, the authors of (Lu et al., 2019) also utilised the information across different tasks which lie at the intersection of vision and natural language processing to improve the performance of their model across all tasks. The authors of (Rawat et al., 2019) utilised weak supervision to the model while predicting the answer but not much work has been done to incorporate the logical form of the question for unstructured question answering in a multi-task setting. Hence, we decided to explore this direction and incorporate the structured semantic information of the questions for extractive question answering.

8 Conclusion

The proposed entity-enriched QA models trained with an auxiliary task improve over the state-of-the-art models by about 3 – 6% across the large-scale clinical QA dataset, *emrQA* (Pampari et al., 2018) (as well as MADE (Jagannatha et al., 2019)). We

also show that multitask learning for logical forms along with the answer results in better generalizing over unseen paraphrases for EMR QA. The predicted logical forms also serve as an accompanying justification to the answer and help in adding credibility to the predicted answer for the physician.

Acknowledgement

This work is supported by MIT-IBM Watson AI Lab, Cambridge, MA USA.

References

- Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *NAACL-HLT Clinical NLP Workshop*.
- Alan R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *AMIA*, page 17.
- Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544.
- Rahul Bhagat, Eduard Hovy, and Siddharth Patwardhan. 2009. Acquiring paraphrases from text corpora. In *Proceedings of the fifth international conference on Knowledge capture*, pages 161–168.
- Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.
- Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075*.
- Qingqing Cai and Alexander Yates. 2013. Large-scale semantic parsing via schema matching and lexicon extension. In *ACL*, volume 1, pages 423–433.
- Deborah A Dahl, Madeleine Bates, Michael Brown, William Fisher, Kate Hunicke-Smith, David Pallett, Christine Pao, Alexander Rudnick, and Elizabeth Shriberg. 1994. Expanding the scope of the atis task: The atis-3 corpus. In *HLT*, pages 43–48.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*.
- Abhyuday Jagannatha, Feifan Liu, Weisong Liu, and Hong Yu. 2019. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (made 1.0). *Drug safety*, 42(1):99–111.
- Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.
- Simon Liu, Wei Ma, Robin Moore, Vikraman Ganesan, and Stuart Nelson. 2005. RxNorm: prescription for electronic drug information exchange. *IT professional*, 7(5):17–23.
- Vanessa Lopez, Christina Unger, Philipp Cimiano, and Enrico Motta. 2013. Evaluating question answering over linked data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 21:3–13.
- Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2019. 12-in-1: Multi-task vision and language representation learning. *arXiv preprint arXiv:1912.02315*.
- Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.
- Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.
- Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*.
- Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrqa: A large corpus for question answering on electronic medical records. *EMNLP*.
- Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don’t know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.
- Bhanu Pratap Singh Rawat, Fei Li, and Hong Yu. 2019. Naranjo question answering using end-to-end multi-task learning model. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2547–2555. ACM.
- Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

- Kent A Spackman, Keith E Campbell, and Roger A Côté. 1997. Snomed rt: a reference terminology for health care. In *Proceedings of the AMIA annual fall symposium*, page 640. American Medical Informatics Association.
- Simon Šuster and Walter Daelemans. 2018. Clicr: a dataset of clinical case reports for machine reading comprehension. *arXiv preprint arXiv:1803.09720*.
- Wilson L Taylor. 1953. “cloze procedure”: A new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433.
- George Tsatsaronis, Michael Schroeder, Georgios Paliouras, Yannis Almirantis, Ion Androutsopoulos, Eric Gaussier, Patrick Gallinari, Thierry Artieres, Michael R Alvers, Matthias Zschunke, et al. 2012. Bioasq: A challenge on large-scale biomedical semantic indexing and question answering. In *2012 AAAI Fall Symposium Series*.
- Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Yushi Wang, Jonathan Berant, and Percy Liang. 2015. Building a semantic parser overnight. In *ACL-IJCNLP*, volume 1, pages 1332–1342.
- Wei-Hung Weng, Yuannan Cai, Angela Lin, Fraser Tan, and Po-Hsuan Cameron Chen. 2019. Multi-modal multitask representation learning for pathology biobank metadata prediction. *arXiv preprint arXiv:1909.07846*.
- Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.
- John M Zelle and Raymond J Mooney. 1996. Learning to parse database queries using inductive logic programming. In *Proceedings of the national conference on artificial intelligence*, pages 1050–1055.
- Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019a. Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*.
- Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, and Hai Zhao. 2019b. Sg-net: Syntax-guided machine reading comprehension. *arXiv preprint arXiv:1908.05147*.
- Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Model Hyper-parameters

Most of the hyper-parameters across our models remained same: learning rate: $2e - 5$, weight decay: $1e - 5$, warm-up proportion: 10% and hidden dropout probability: 0.1. The parameters that varied across models for different datasets are enumerated in the Table 6. The hyper-parameters provided in Table 6 are for all models in a particular dataset. This also suggests that even after adding an auxiliary task, the proposed model doesn't need a lot of hyper-parameter tuning.

Dataset	Entity Embedding Dim	Auxiliary Task Wt.
emrQA-rel	100	0.3
BoolQ	90	0.3
emrQA	100	0.3
MADE	80	0.2

Table 6: Hyper-parameter values across different datasets.

B Logical forms (LFs) for MADE dataset

1. MedicationEvent ($|medication|$) [sig=x]
2. MedicationEvent ($|medication|$) causes ConditionEvent (x) OR SymptomEvent (x)
3. MedicationEvent ($|medication|$) given ConditionEvent (x) OR SymptomEvent (x)
4. [ProcedureEvent ($|treatment|$) given/conducted ConditionEvent (x) OR SymptomEvent (x)] OR [MedicationEvent ($|treatment|$) given ConditionEvent (x) OR SymptomEvent (x)]
5. MedicationEvent (x) CheckIfNull ([enddate]) OR MedicationEvent (x) [enddate>currentDate] OR ProcedureEvent (x) [date=x] given ConditionEvent ($|problem|$) OR SymptomEvent ($|problem|$)
6. MedicationEvent (x) CheckIfNull ([enddate]) OR MedicationEvent (x) [enddate>currentDate] given ConditionEvent ($|problem|$) OR SymptomEvent ($|problem|$)
7. MedicationEvent ($|treatment|$) OR ProcedureEvent ($|treatment|$) given ConditionEvent (x) OR SymptomEvent (x)
8. MedicationEvent ($|treatment|$) OR ProcedureEvent ($|treatment|$) improves/worsens/causes ConditionEvent (x) OR SymptomEvent (x)

C Selected entities from MetaMap

The list of selected semantic types in the form of entities and their brief descriptors are provided in Table 7.

Semantic Type	Description
acab	Acquired Abnormality
aggp	Age Group
anab	Anatomical Abnormality
anst	Anatomical Structure
bpoc	Body Part, Organ, or Organ Component
cgab	Congenital Abnormality
clnd	Clinical Drug
diap	Diagnostic Procedure
emod	Experimental Model of Disease
evnt	Event
findg	Finding
inpo	Injury or Poisoning
lbpr	Laboratory Procedure
lbtr	Laboratory or Test Result
phob	Physical Object
qnco	Quantitative Concept
sbst	Substance
sosy	Sign or Symptom
topp	Therapeutic or Preventive Procedure

Table 7: Selected semantic types as per MetaMap and their brief descriptions.