IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 1, 2019, accepted April 23, 2019, date of publication April 29, 2019, date of current version May 14, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2913694

Incorporating Domain Knowledge into Natural Language Inference on Clinical Texts

MINGMING LU[1], YU FANG[1], FENGQI YAN[1], AND MAOZHEN LI[2]
[1]Department of Computer Science and Technology, Tongji University, Shanghai 201804, China
[2]Department of Electronic and Computer Engineering, Brunel University London, Uxbridge, UB8 3PH, U.K.

Corresponding author: Yu Fang (fangyu@tongji.edu.cn)

This work was supported by the Fundamental Research Funds for the Central Universities under Grant 22120180117.

ABSTRACT Making inference on clinical texts is a task which has not been fully studied. With the newly released, expert annotated MedNLI dataset, this task is being boosted. Compared with open domain data, clinical texts present unique linguistic phenomena, e.g., a large number of medical terms and abbreviations, different written forms for the same medical concept, which make inference much harder. Incorporating domain-specific knowledge is a way to eliminate this problem, in this paper, we assemble a new *incorporating medical concept definitions* module on the classic enhanced sequential inference model (ESIM), which first extracts the most relevant medical concept for each word, if it exists, then encodes the definition of this medical concept with a bidirectional long short-term network (BiLSTM) to obtain domain-specific definition representations, and attends these definition representations over vanilla word embeddings. The empirical evaluations are conducted to demonstrate that our model improves the prediction performance and achieves a high level of accuracy on the MedNLI dataset. Specifically, the knowledge enhanced word representations contribute significantly to entailment class.

INDEX TERMS Attention mechanism, clinical text, medical domain knowledge, natural language inference, word representation.

I. INTRODUCTION

Natural Language Inference (NLI), also known as Recognizing Textual Entailment (RTE), is a task concerning semantic relationship (*entailment*, *contradiction*, or *neutral*) between a *premise* and a *hypothesis* [1]. In recent years, represented by the Stanford Natural Language Inference (SNLI) [2] corpus and the Multi-Genre Natural Language Inference (MultiNLI) [3] corpus, large-scale annotated datasets are made publicly available, which have pushed the development of this task. In addition, many deep neural network models are proposed to achieve the state-of-the-art performance [4]–[6].

In the clinical domain, newly released MedNLI [7] dataset focuses on NLI task on clinical texts. Owing to the specialty and particularity of this domain, clinical texts present unique linguistic phenomena different from open domain data: (1) the existence of a large number of medical terms and abbreviations leads to the out-of-vocabulary (OOV) issue; (2) a medical concept has different written forms in different vocabularies, though they have the same meaning. Table 1 are some examples from the MedNLI dataset for illustration. The key words in Example #1 are "*diaphoresis*" and "*sweats*", which express the same medical concept, but they are written in different forms. Example #2 and #3 have medical terms ("*lumbar puncture*" and "*coronary artery bypass grafting*"), as well as standard medical abbreviations ("*LP*" and "*STEMI*") and not standard logogram words ("*pt*", meaning patient). If a system cannot understand these medical terms and abbreviations correctly, it will misclassify the classes. In general, these unique linguistic phenomena make inference on MedNLI much harder.

Since processing of clinical texts requires domain-specific knowledge, in this paper, we incorporate such knowledge into the classic open domain model (ESIM) by encoding the definitions of medical concepts with a bidirectional LSTM [8] (BiLSTM) and attending the vanilla word embeddings to these domain-specific representations. Through this way, computers are taught to, on one hand, learn the meanings of medical terms and abbreviations, on the other hand,

The associate editor coordinating the review of this manuscript and approving it for publication was Yonghong Peng.

TABLE 1. Examples from the MedNLI dataset. P, H, and L stand for premise, hypothesis, and label, respectively. Domain-specific words for inference are in italics. "*LP*" is the abbreviation for lumbar puncture and "*STEMI*" stands for ST segment elevation myocardial infarction.

Example #1 **P**: It was also associated with *diaphoresis*. **H**: The patient has *sweats*. **L**: entailment
Example #2 **P**: The pt was transfered to have an *LP* with neurosurg backup. **H**: The patient has no neurological symptoms, or indication for *lumbar puncture* **L**: contradiction
Example #3 **P**: He presented preoperatively for *coronary artery bypass grafting*. **H**: Patient has had a *STEMI* **L**: neutral

identify similarities and differences between medical concepts. We conduct experiments on the MedNLI dataset, and the results showing that our model outperforms all baselines done by Romanov and Shivade [7], achieving the state-of-the-art performance. In addition, we present ablation study and case study to learn how domain knowledge contributes to our model.

Our work has three main contributions:

- We propose a knowledge enhanced model for natural language inference on clinical texts, which combines BiLSTM and attention to enhance vanilla word embeddings with definitions of medical concepts.
- We study of the effectiveness of our model on the MedNLI dataset, and achieve a higher level of accuracy than those models without knowledge enhanced.
- Our ablation study and case study reveal some useful insights for the contributions of knowledge enhanced word representations.

The rest of this paper is organized as follows. Section II reviews the related work for natural language inference. Section III details the design of the proposed model. Section IV and V present and discuss the experimental settings and results, respectively. Finally, we draw conclusion in Section VI.

II. RELATED WORK

There are two types of approaches for natural language inference task: **encoding-based models** and **interaction-based models** [9]. Encoding-based models [2], [4], [10], [11] use siamese architecture [12] to learn vector representations of the premise and hypothesis, and then calculate the semantic relationship between two sentences based on a neural network classifier. One representative model is InferSent [4], which is one baseline model of the MedNLI dataset.

Interaction-based models [5], [13], [14] utilize some sorts of word alignment mechanisms, e.g., attention [15], then aggregate inter-sentence interactions. As shown in the SemEval-2016 task of interpretable semantic textual similarity [16], the semantic relations of aligned chunks contribute a lot to sentence pair modeling, interaction-based models have better performance than encoding-based models. Chen *et al.* [5] proposed an enhanced sequential inference model (ESIM), which contains three main components, i.e., input encoding, co-attention matching, and inference composition. ESIM is another baseline model of the MedNLI dataset.

Unlike previous work [6] that enriches NLI models with lexical-level semantic knowledge about synonymy, antonymy, hypernymy, hyponymy and co-hyponymy between words, we focus on medical domain and explore the incorporation of extra knowledge on clinical texts for natural language inference. Romanov and Shivade [7] also studied two ways of incorporating domain-specific knowledge into their baseline models. In one way, they modified pre-trained word embeddings by retrofitting [17], so the input to models could carry clinical information. However, this way only degrades the performance. Because retrofitting works only on directly related concepts, while medical concepts are more complex, and medical inferences require more steps of reasoning. Another way is knowledge-directed attention, which is beneficial to the InferSent and ESIM models. Our model is similar to the first way, modifying model's inputs, but we utilize definition representations to enhance the word embeddings of medical terms and abbreviations, alleviating the OOV issue and bridging the semantic gap between different written forms of a medical concept.

III. MODEL DESIGN

In this section, we will explain the NLI task and describe our domain knowledge, i.e., definitions of medical concepts. Then, we study how to incorporate these definitions into the ESIM model for natural language inference on clinical texts.

A. PROBLEM DEFINITION

Given the MedNLI dataset $\mathcal{D}$, an example of the dataset can be represented as a (p, h, y) triplet consisting of premise p, hypothesis h, and ground truth label y. Specially, the premise is represented as $p = \{a_i\}_{i=1}^{M}$ and the hypothesis is $h = \{b_j\}_{j=1}^{N}$, where M and N are the lengths of the sentences. $y \in \{0, 1, 2\}$ is the corresponding label of the given triple which takes a value of 0 if the premise entails the hypothesis (*entailment*), 1 if they contradict each other (*contradiction*), and 2 if they are unrelated (*neutral*). Our goal is to learn a predictive distribution $p(y|p, h; \boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta}$ from $\mathcal{D}$. That is, given a premise p and hypothesis h, we would like to infer the probability that they will be classified as entailment, contradiction, or neutral.

B. DOMAIN KNOWLEDGE

First, we collect the definitions of medical concepts from Unified Medical Language System (UMLS) [18]. In UMLS, for a medical concept, there would be multiple definitions coming from different source vocabularies. To simplify the model, we choose the shortest one as the only definition

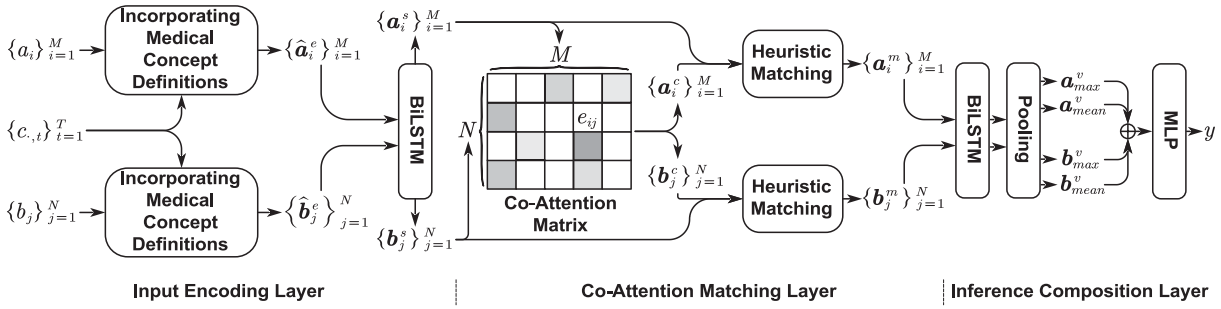

FIGURE 1. An overview of our model. Similar to the ESIM model, our model consists of three layers, i.e., input encoding layer, co-attention matching layer, and inference composition layer. The difference is that we incorporate medical concept definitions in the first layer. $\{a_i\}_{i=1}^{M}$, $\{b_j\}_{j=1}^{N}$, and $\{c_{\cdot,t}\}_{t=1}^{T}$ are the inputs to the model, representing the premise sentence, hypothesis sentence, and the definitions of extracted medical concepts from two sentence, respectively. y is the output. $\oplus$ means concatenation of vectors.

TABLE 2. Some examples of medical concepts and their definitions from UMLS.

Word	Medical Concept	Definition
diaphoresis	Increased sweating	Profuse sweating.
LP	Spinal Puncture	Tapping fluid from the subarachnoid space in the lumbar region, usually between the third and fourth lumbar vertebrae.
STEMI	ST segment elevation myocardial infarction	A clinical syndrome defined by MYOCARDIAL ISCHEMIA symptoms; persistent elevation in the ST segments of the ELECTROCARDIOGRAM; and release of BIOMARKERS of myocardia NECROSIS (e.g., elevated TROPONIN levels).

of this medical concept. In the end, we collect a total of 198,042 definitions that make up our domain knowledge base, denoted as $\mathcal{K}$.

Second, following the previous work [7], we use Metamap [19] to extract medical concepts from premise and hypothesis sentences, and map them to standard terminologies in the UMLS. For each extracted phrase, there may be more than one related concepts, which are sorted by MetaMap Indexing (MMI) score. The higher the score, the greater the relevance of the medical concept to its extracted phrase. In this paper, we only consider the concept with the highest score for each word, and discard those with the lower scores. As a result, every word has zero or one corresponding medical concept. Through this way, we know exactly what concept the medical term or abbreviation stands for, and different written forms could be mapped to the same concept. Finally, we associate words with concept definitions. For example, if one word a_i in the premise sentence extracts a medical concept, then we search our domain knowledge base $\mathcal{K}$ for its definition. We denote word a_i associated definition as $\{c_{i,t}\}_{t=1}^{T}$. Table 2 shows the extracted medical concepts of some domain-specific words of Example #1 to #3, and their definitions from UMLS.

C. MODEL OVERVIEW

We present here our model for natural language inference on clinical texts. It consists of three layers: input encoding layer, co-attention layer, and inference composition layer. Fig. 1 shows an overview of our model.

The model takes the premise sentence, the hypothesis sentence, and the definitions of extracted medical concepts from two sentences as inputs, and then first constructs respective word representations with pre-trained word embeddings. These pre-trained word embeddings can be either publicly available open domain word embeddings, or trained on a domain-specific corpora. Then, each word in two sentences are attended over their corresponding definition if it exists, which is done by the *Incorporating Medical Concept Definitions* module. Furthermore, the enhanced word embeddings are fed into a siamese BiLSTM network to obtain a set of contextualized representations of premise and hypothesis sentences.

In the co-attention matching layer, we use soft-alignment of contextualized word representations between the premise and hypothesis to obtain aligned representation, followed by a heuristic matching approach [20] to collect local inference vectors for each word. Finally, to determine the overall inference relationship between the premise and hypothesis, another BiLSTM is utilized to compose the collected local inference vectors, which is part of the inference composition layer. The output hidden vectors of the second BiLSTM are converted to fixed-length vectors with max and mean pooling operations and put into the final multilayer perceptron (MLP) classifier to determine the inference class.

Details about each layer and the *Incorporating Medical Concept Definitions* module are provided in the following sections.

D. INPUT ENCODING LAYER

Input encoding layer takes as inputs the premise $\{a_i\}_{i=1}^{M}$, the hypothesis $\{b_j\}_{j=1}^{N}$, and associated medical concept definitions $\{c_{\cdot,t}\}_{t=1}^{T}$, where $\cdot$ can be replaced with i or j. Pre-trained word embeddings $\boldsymbol{E} \in \mathbb{R}^{d_e \times |V|}$ are first used to converted word inputs to vector sequences $\boldsymbol{a}_1^e, \ldots, \boldsymbol{a}_M^e]$, $[\boldsymbol{b}_1^e, \ldots, \boldsymbol{b}_M^e]$, and $[\boldsymbol{c}_{\cdot,1}^e, \ldots, \boldsymbol{c}_{\cdot,T}^e]$, where $|V|$ is the vocabulary size and d_e is the dimension of the word embedding. In the experiments, we explore six different word embeddings, one publicly available open domain word embedding, two trained on domain-specific corpus, and three initialized with open domain word embeddings and further fine-tuned on one or two domain-specific corpus:

- **GloVe**$_{[\text{CC}]}$: GloVe embeddings [21], trained on Common Crawl.
- **fastText**$_{[\text{BioASQ}]}$: fastText embeddings [22], trained on PubMed abstracts from the BioASQ challenge [23].
- **fastText**$_{[\text{MIMIC-III}]}$: fastText embeddings, trained on patient clinical notes from the MIMIC-III database [24].
- **GloVe**$_{[\text{CC}]} \rightarrow$ **fastText**$_{[\text{BioASQ}]}$: GloVe embeddings for initialization and further fine-tuned on the BioASQ data.
- **GloVe**$_{[\text{CC}]} \rightarrow$ **fastText**$_{[\text{BioASQ}]} \rightarrow$ **fastText**$_{[\text{MIMIC-III}]}$: GloVe embeddings for initialization and further fine-tuned on the BioASQ and MIMIC-III data in succession.
- **fastText**$_{[Wiki]} \rightarrow$,**fastText**$_{[\text{MIMIC-III}]}$: fastText Wikipedia embeddings for initialization and further fine-tuned on the MIMIC-III data.

All of the domain-specific word embeddings are downloaded from the MedNLI dataset.[1]

1) INCORPORATING MEDICAL CONCEPT DEFINITIONS

Inspired by the work of [25] and [26], we incorporate medical concept definitions into word embeddings, as shown in Fig. 2.

The bidirectional long short-term memory (BiLSTM) network has been proven to be good at modeling dependencies coming from both the past and the future in sequences. So we employ it to encode definition embeddings in forward and backward directions. Take Fig. 2 for example, $\boldsymbol{c}_t^e$ is the input to the BiLSTM at time step t. To simplify notation, we omit the subscript i in this section. The hidden states in the forward direction are updated as follows:

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}^i \boldsymbol{c}_t^e + \boldsymbol{U}^i \overrightarrow{\boldsymbol{h}}_{t-1} + \boldsymbol{b}^i) \quad (1)$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}^f \boldsymbol{c}_t^e + \boldsymbol{U}^f \overrightarrow{\boldsymbol{h}}_{t-1} + \boldsymbol{b}^f) \quad (2)$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}^o \boldsymbol{c}_t^e + \boldsymbol{U}^o \overrightarrow{\boldsymbol{h}}_{t-1} + \boldsymbol{b}^o) \quad (3)$$

$$\boldsymbol{q}_t = \tanh(\boldsymbol{W}^q \boldsymbol{c}_t^e + \boldsymbol{U}^q \overrightarrow{\boldsymbol{h}}_{t-1} + \boldsymbol{b}^q) \quad (4)$$

$$\boldsymbol{p}_t = \boldsymbol{f}_{i,t} \circ \boldsymbol{p}_{i,t-1} + \boldsymbol{i}_t \circ \boldsymbol{q}_t \quad (5)$$

$$\overrightarrow{\boldsymbol{h}}_t = \boldsymbol{o}_t \circ \tanh(\boldsymbol{p}_t) \quad (6)$$

where $\boldsymbol{i}_t$, $\boldsymbol{f}_t$, $\boldsymbol{o}_t$ are the input gate, forget gate and output gate of LSTM, respectively. σ is the sigmoid function, and $\boldsymbol{p}_t$ is the cell state. Accordingly, in the forward direction,

[1] https://jgc128.github.io/mednli/

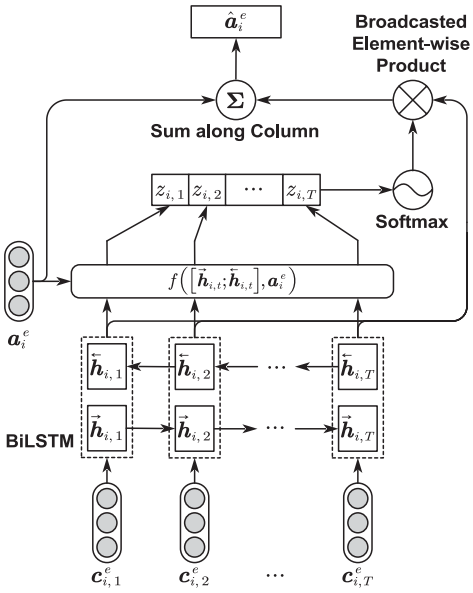

FIGURE 2. An illustration of incorporating medical concept definition embeddings $\{c_{i,t}^e\}_{t=1}^{T}$ into the vanilla word embedding a_i^e of one medical term or abbreviation in the premise. From the output, we will get the enhanced word representation $\hat{a}_i^e$.

the hidden state $\overrightarrow{\boldsymbol{h}}_t$ at time step t depends on input word and the preceding hidden state $\overrightarrow{\boldsymbol{h}}_{t-1}$. Similarly, in the backward direction, the hidden state $\overleftarrow{\boldsymbol{h}}_t$ is updated based on current input and the hidden state from the next time step. At the t-th time step, the output of BiLSTM is usually obtained by concatenation of the hidden states from both directions, formally, $\boldsymbol{h}_t = [\overrightarrow{\boldsymbol{h}}_t; \overleftarrow{\boldsymbol{h}}_t]$. Especially, the above process can be simplified as a BiLSTM function:

$$\boldsymbol{h}_1, \ldots, \boldsymbol{h}_T = \text{BiLSTM}(\boldsymbol{c}_1^e, \ldots, \boldsymbol{c}_T^e) \quad (7)$$

To obtain definition enhanced word embeddings, we utilize a multi-layer perceptron attention [15] mechanism to aggregate the outputs of BiLSTM and then add them to the vanilla word embeddings. In particular, attention first computes the alignment score between $\boldsymbol{h}_t$ and $\boldsymbol{a}^e$ by a function $f(\boldsymbol{h}_t, \boldsymbol{a}^e)$:

$$f(\boldsymbol{h}_t, \boldsymbol{a}^e) = \boldsymbol{v}^T \sigma(\boldsymbol{W}^h \boldsymbol{h}_t + \boldsymbol{W}^e \boldsymbol{a}^e) \quad (8)$$

where $\boldsymbol{W}^h$, $\boldsymbol{W}^e$ are weight matrices and $\boldsymbol{v}$ is a weight vector. This alignment score measures the attention of $\boldsymbol{a}^e$ to $\boldsymbol{h}_t$. Subsequently, a softmax function normalizes alignment scores to form a vector $\boldsymbol{z} \in \mathbb{R}^T$:

$$z_t = \frac{\exp(f(\boldsymbol{h}_t, \boldsymbol{a}^e))}{\sum_{t'=1}^{T} \exp(f(\boldsymbol{h}_{t'}, \boldsymbol{a}^e))} \quad (9)$$

Here, z_t is an indicator of the importance of $\boldsymbol{h}_t$ to $\boldsymbol{a}^e$. So, the output of attention is a weighted sum of $\{\boldsymbol{h}_t\}_{t=1}^{T}$, where the weights are given by $\boldsymbol{z}$.

By adding the output of attention and the vanilla word embedding, we obtain definition enhanced word embedding

in the premise:

$$\hat{\boldsymbol{a}}^e = \sum_{t=1}^{T} z_t \boldsymbol{h}_t + \boldsymbol{a}^e \tag{10}$$

The above approach of incorporating medical concept definitions also applies to the hypothesis.

2) SENTENCE ENCODING

To represent words in their context, the enhanced word embeddings of premise and hypothesis are fed into a parameters shared BiLSTM to obtain contextualized representations $\boldsymbol{a}^s$ and $\boldsymbol{b}^s$:

$$\boldsymbol{a}^s_1, \ldots, \boldsymbol{a}^s_M = \text{BiLSTM}_1(\hat{\boldsymbol{a}}^e_1, \ldots, \hat{\boldsymbol{a}}^e_M) \tag{11}$$
$$\boldsymbol{b}^s_1, \ldots, \boldsymbol{b}^s_N = \text{BiLSTM}_1(\hat{\boldsymbol{b}}^e_1, \ldots, \hat{\boldsymbol{b}}^e_N) \tag{12}$$

E. CO-ATTENTION MATCHING LAYER

Modeling the interactions is the critical component for deciding the inference relationship between the premise and hypothesis. In this layer, a co-attention matrix is computed using dot-product to produce aligned word representations, and then by comparing with contextualized representations, we collect matching information at the word level.

First, the co-attention score between each representation tuple $(\boldsymbol{a}^s_i, \boldsymbol{b}^s_j)$ is calculated as follows:

$$e_{ij} = (\boldsymbol{a}^s_i)^T \boldsymbol{b}^s_j \tag{13}$$

Then for the i-th word in the premise, its relevant representation carried by the hypothesis is identified and composed using e_{ij} as

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{j'=1}^{N} \exp(e_{ij'})} \tag{14}$$
$$\boldsymbol{a}^c_i = \sum_{j=1}^{N} \alpha_{ij} \boldsymbol{b}^s_j \tag{15}$$

where $\alpha \in \mathbb{R}^{M \times N}$ is the normalized co-attention matrix w.r.t. the column-axis, and $\boldsymbol{a}^c_i$ is a weighted sum of $\{\boldsymbol{b}^s_j\}_{j=1}^N$, meaning the contents related to $\boldsymbol{a}^s_i$ are selected to form $\boldsymbol{a}^c_i$. The same calculation is performed for each word in the hypothesis as

$$\beta_{ij} = \frac{\exp(e_{ij})}{\sum_{i'=1}^{M} \exp(e_{i'j})} \tag{16}$$
$$\boldsymbol{b}^c_j = \sum_{i=1}^{M} \beta_{ij} \boldsymbol{a}^s_i \tag{17}$$

where $\beta \in \mathbb{R}^{M \times N}$ is the normalized co-attention matrix w.r.t. the row-axis. We denote $\boldsymbol{a}^c_i$ and $\boldsymbol{b}^c_j$ as aligned word representations.

To further enhance inference information, followed the heuristic matching approach proposed by Mou *et al.* [20], we concatenate contextualized and aligned word representations with the differences and element-wise products between each other, resulting local inference vectors. Formally, local inference vectors $\boldsymbol{a}^m_i$ and $\boldsymbol{b}^m_j$ are calculated as follows:

$$\boldsymbol{a}^m_i = G([\boldsymbol{a}^s_i; \boldsymbol{a}^c_i; \boldsymbol{a}^s_i - \boldsymbol{a}^c_i; \boldsymbol{a}^s_i \circ \boldsymbol{a}^c_i]) \tag{18}$$
$$\boldsymbol{b}^m_j = G([\boldsymbol{b}^s_j; \boldsymbol{b}^c_j; \boldsymbol{b}^s_j - \boldsymbol{b}^c_j; \boldsymbol{b}^s_j \circ \boldsymbol{b}^c_j]) \tag{19}$$

where G is one-layer feed-forward neural network with the ReLU [27] activation function to reduce dimensionality.

F. INFERENCE COMPOSITION LAYER

In this layer, a parameters shared BiLSTM followed by max and mean pooling operations is typically employed as the aggregation method to compose the local inference vectors collected above:

$$\boldsymbol{a}^v_1, \ldots, \boldsymbol{a}^v_M = \text{BiLSTM}_2(\boldsymbol{a}^m_1, \ldots, \boldsymbol{a}^m_M) \tag{20}$$
$$\boldsymbol{b}^v_1, \ldots, \boldsymbol{b}^v_N = \text{BiLSTM}_2(\boldsymbol{a}^m_1, \ldots, \boldsymbol{b}^m_N) \tag{21}$$
$$\boldsymbol{a}^v_{max} = \max_{1 \leqslant i \leqslant M} \boldsymbol{a}^v_i \tag{22}$$
$$\boldsymbol{a}^v_{mean} = \text{mean}_{1 \leqslant i \leqslant M} \boldsymbol{a}^v_i \tag{23}$$
$$\boldsymbol{b}^v_{max} = \max_{1 \leqslant j \leqslant N} \boldsymbol{b}^v_j \tag{24}$$
$$\boldsymbol{b}^v_{mean} = \text{mean}_{1 \leqslant j \leqslant N} \boldsymbol{b}^v_j \tag{25}$$

Again we use BiLSTM here, but the role is completely different from that presented in Section III-D.2. The BiLSTM here learns to discriminate critical local inference vectors for obtaining the overall sentence-level inference relationship between the premise and hypothesis. The pooling vectors are concatenated together and fed into the final multi-layer perceptron (MLP) classifier which has one hidden layer with tanh activation and *softmax* output layer:

$$\boldsymbol{y} = \text{MLP}([\boldsymbol{a}^v_{max}; \boldsymbol{a}^v_{mean}; \boldsymbol{b}^v_{max}; \boldsymbol{b}^v_{mean}]) \tag{26}$$

where $\boldsymbol{y} \in \mathbb{R}^3$, and each entry is the probability distribution $p(y|p, h)$ over class y.

G. OPTIMIZATION OBJECTIVE

The entire model is trained in an end-to-end manner via minimizing the multi-class cross-entropy loss. The loss function is defined as:

$$J(\boldsymbol{\theta}) = -\frac{1}{|\mathcal{D}|} \sum_i \log(p(\hat{y}_i|p_i, h_i)) \tag{27}$$

where $\boldsymbol{\theta}$ denotes all trainable parameters, $|\mathcal{D}|$ is the number of training examples, and $\hat{y}_i$ is the ground truth for the i-th example.

IV. EXPERIMENTS

In this section, we first briefly introduce the MedNLI dataset, a newly released dataset for natural language inference on clinical texts, followed by detailed training settings.

A. MEDNLI DATASET

We evaluated our model on the MedNLI dataset [7], which contains 13k expert annotated sentence pairs. The premise

TABLE 3. Accuracies of our model (ESIM w/ Knowledge) compared to baselines using different word embeddings on MedNLI. Baseline results are directly copied from Romanov and Shivade [7].

Word Embeddings	InferSent Baselines	ESIM Baselines	ESIM w/ Knowledge
$\text{GloVe}_{[\text{CC}]}$	0.735	0.731	0.742
$\text{fastText}_{[\text{BioASQ}]}$	0.741	0.733	0.753
$\text{fastText}_{[\text{MIMIC-III}]}$	0.758	0.743	**0.778**
$\text{GloVe}_{[\text{CC}]} \rightarrow \text{fastText}_{[\text{BioASQ}]}$	0.742	0.745	0.765
$\text{GloVe}_{[\text{CC}]} \rightarrow \text{fastText}_{[\text{BioASQ}]} \rightarrow \text{fastText}_{[\text{MIMIC-III}]}$	0.762	0.749	0.776
$\text{fastText}_{[\text{Wiki}]} \rightarrow \text{fastText}_{[\text{MIMIC-III}]}$	**0.766**	0.748	0.771

sentences were drawn from clinical notes contained in the MIMIC-III v1.3 database [24], and the hypothesis sentences were generated by four clinicians. The resulting dataset consists of 14,049 pairs of premises and hypotheses. Among them, there are 11,232 pairs for training, 1,395 pairs for development, and 1,422 pairs for testing. The average sentence lengths of premises and hypotheses are 20 and 5.8 respectively. Meanwhile, the maximum sentence lengths of premises and hypotheses are 202 and 20 respectively. We use the same data split as provided in Romanov and Shivade [7] and classification accuracy as our evaluation metric.

B. TRAINING DETAILS

Following all baselines' settings on the MedNLI dataset, we chose the dimension of word embeddings and hidden states of BiLSTMs of 300, except for the BiLSTM in the *Incorporating Medical Concept Definitions* module, which was 150. We restricted the lengths of the premise and hypothesis sentences by a maximum of 50 words, and that of medical concept definitions by 200. All word embeddings were fixed during training. Adam [28] was used for optimization with an initial learning rate of 0.001. The mini-batch size was set to 64. We set a dropout rate of 0.5 for input and output of hidden layer of the final MLP classifier. We also used variational dropout [29] for input of BiLSTMs, which was also set to 0.5. We trained our model for a maximum of 20 epochs. The training was stopped when the development loss did not decrease after 5 subsequent epochs.

All hyper-parameters were strictly selected on the development set, and then tested on the corresponding test set. We used PyTorch[2] and AllenNLP[3] to implement our model.

V. RESULTS

In this section, we will analyze the performance of our model from three aspects. First, we will compare our model with baseline models for different word embeddings. Then, ablation study and case study are conducted to inspect how domain knowledge contributes to the model.

A. COMPARISON AGAINST BASELINES

We compare our model, referred to as **ESIM w/ Knowledge**, against **InferSent** and **ESIM** baseline models tested by Romanov and Shivade [7] for six different word

[2]https://pytorch.org/
[3]https://allennlp.org/

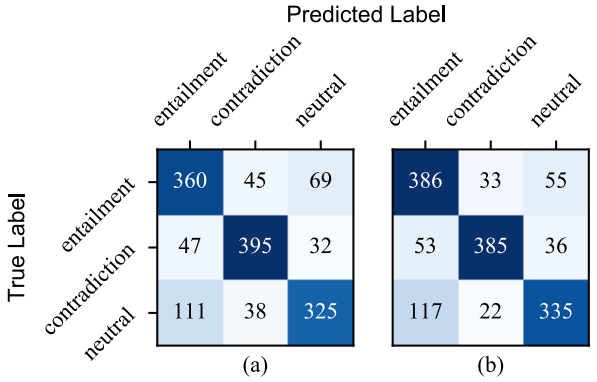

FIGURE 3. Confusion matrix without normalization: (a) InferSent baseline using $\text{fastText}_{[\text{Wiki}]} \rightarrow \text{fastText}_{[\text{MIMIC-III}]}$ embedding[4]; (b) ESIM w/ Knowledge using $\text{fastText}_{[\text{MIMIC-III}]}$ embedding.

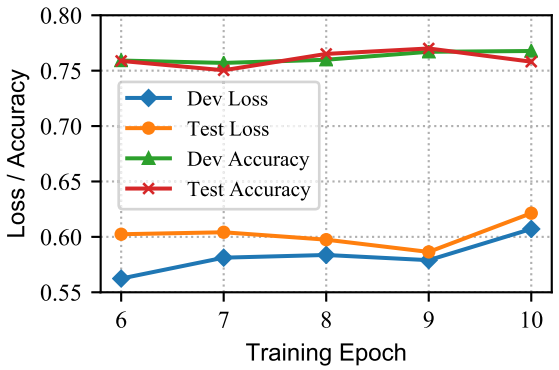

FIGURE 4. Loss and accuracy curve of the development and test set using $\text{fastText}_{[\text{MIMIC-III}]}$ embedding.

embeddings stated in Section III-D. The results are reported in Table 3. Our model outperforms all baseline models and achieves the state-of-the-art performance, indicating that incorporating medical concept definitions can significantly improve the performance. Compared to the best baseline (i.e., InferSent using $\text{fastText}_{[\text{Wiki}]} \rightarrow \text{fastText}_{[\text{MIMIC-III}]}$ embedding), we observed an absolute gain of 0.012 corresponding to 1.6% relative gain in the model using $\text{fastText}_{[\text{MIMIC-III}]}$ embedding. Actually, a total of three results for different word embeddings (others are $\text{GloVe}_{[\text{CC}]} \rightarrow \text{fastText}_{[\text{BioASQ}]} \rightarrow \text{fastText}_{[\text{MIMIC-III}]}$ embedding and $\text{fastText}_{[\text{Wiki}]} \rightarrow \text{fastText}_{[\text{MIMIC-III}]}$ embedding) exceed the best baseline.

In baseline models, all results except one of InferSent are better than those of ESIM. However, for each word embedding, our result goes beyond all two baselines, proving the effectiveness of ESIM integrated with domain

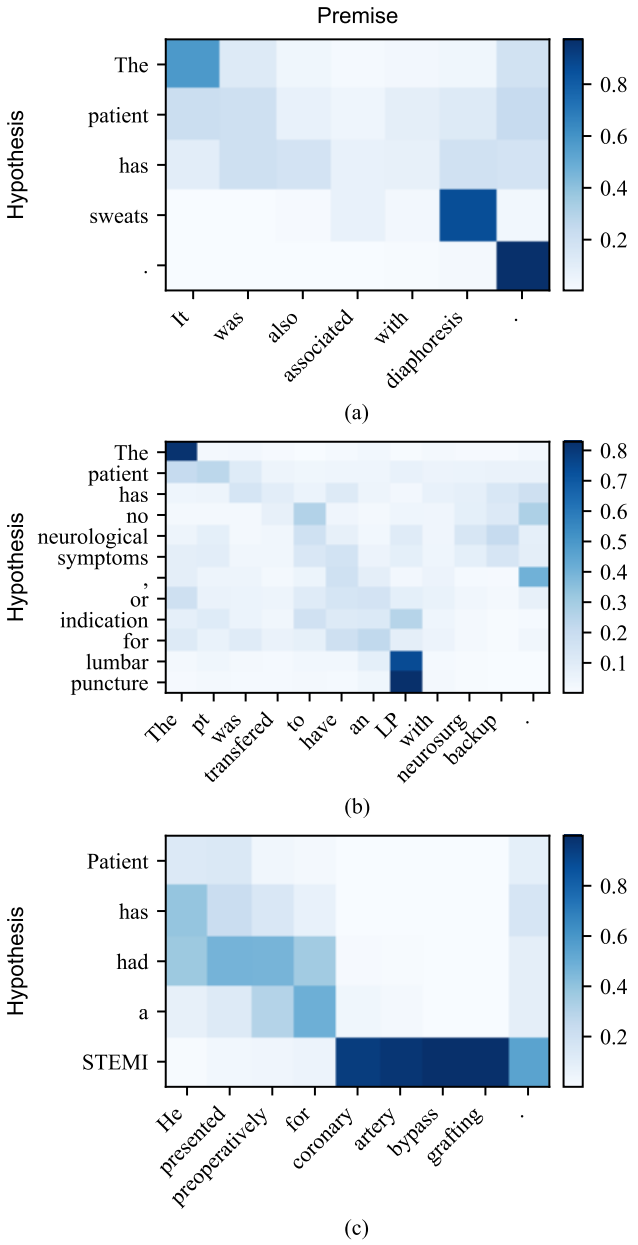

FIGURE 5. Normalized Co-attention matrix of Example #1 to #3. (a) Example #1 with normalization over row-axis. (b) Example #2 with normalization over row-axis. (c) Example #3 with normalization over column-axis.

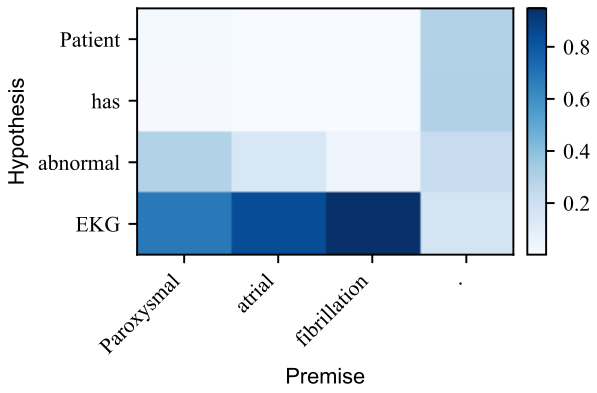

FIGURE 6. Co-attention matrix of Example #4 with normalization over column-axis.

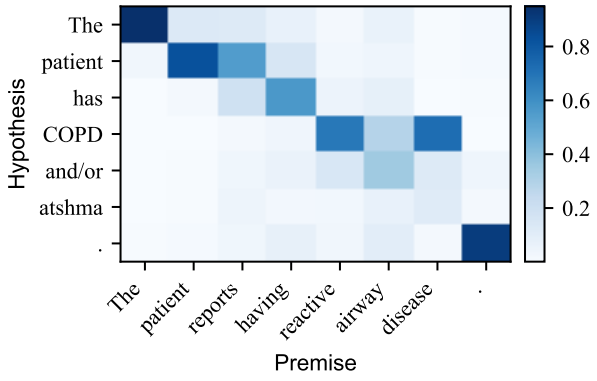

FIGURE 7. Co-attention matrix of Example #5 with normalization over column-axis.

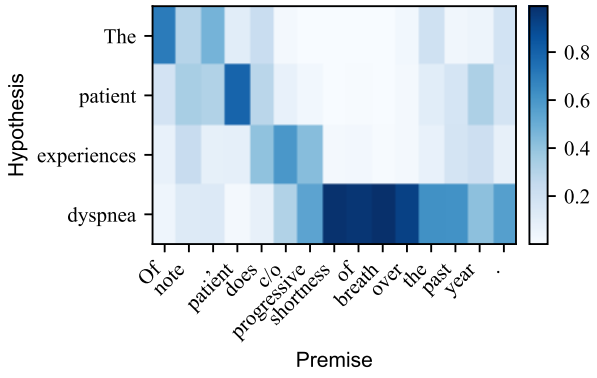

FIGURE 8. Co-attention matrix of Example #6 with normalization over column-axis.

knowledge. The greatest gain of our model is for $\text{GloVe}_{[\text{CC}]} \rightarrow \text{fastText}_{[\text{BioASQ}]}$ embedding (0.765 compared to 0.745), where we obtain an absolute gain of 0.02 and a relative gain of 2.7%.

Besides comparing the overall performance, we also draw the confusion matrix to visualize the classification results of three classes (entailment, contradiction and neutral). As shown in Fig. 3, there are two confusion matrices without normalization, the left belongs to best baseline[4] and the right belongs to the best result of our model. By comparing these two confusion matrices, the following conclusions can be drawn:

(1) Our model improves the performance in entailment and neutral classes, of which it contributes a lot to entailment class, and the misclassifications to contradiction and neural classes are reduced by 12 and 14 respectively. We think this is because the incorporated domain knowledge enhances the word representations of medical terms and abbreviations and bridges the semantic gap between different written forms

[4]Results were predicted by model parameters released by Romanov and Shivade [7], which only obtained an accuracy of 0.759, different from the accuracy of 0.766 stated in the paper.

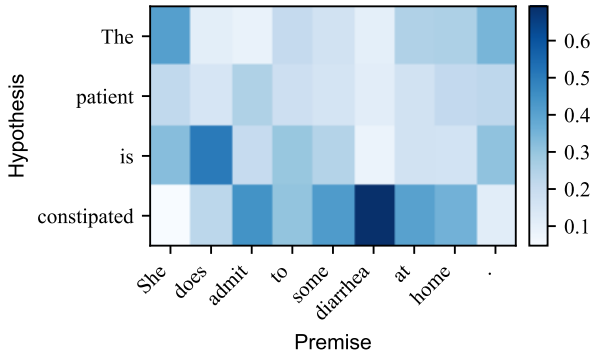

FIGURE 9. Co-attention matrix of Example #7 with normalization over column-axis.

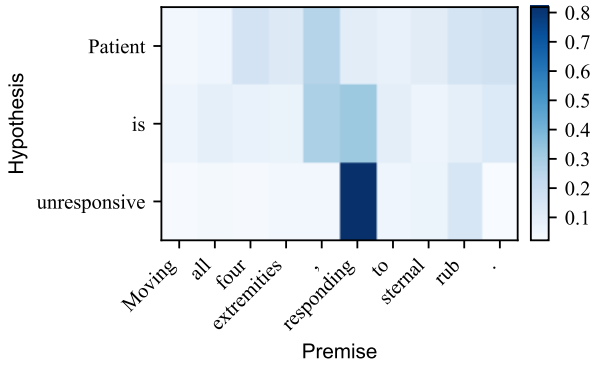

FIGURE 10. Co-attention matrix of Example #8 with normalization over row-axis.

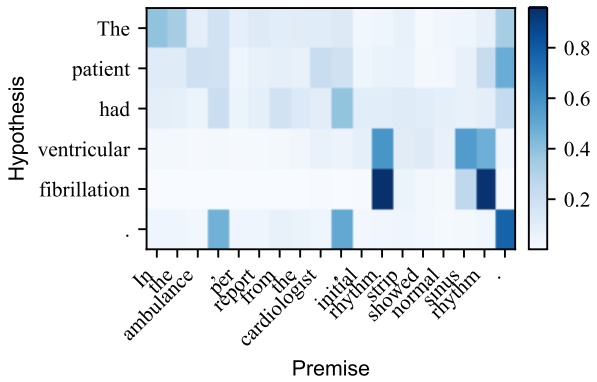

FIGURE 11. Co-attention matrix of Example #9 with normalization over column-axis.

of the same medical concept. The incorporated knowledge also reduces the possibility of neural class being mistakenly classified as contradiction class.

(2) Our model beats the performance in contradiction class. After reviewing the misclassified examples, we found that the errors mainly occurred in those requiring numerical reasoning, e.g., a premise as "*In the ED, initial VS revealed T 98.9, HR 73, BP 121/90, RR 15, O2 sat 98% on RA*." Our model tends to mistake such numerical reasoning examples for entailment class. This is also true in neural class. We think ensemble methods using InferSent and ESIM w/ Knowledge will take advantages of each model and obtain better predictive performance.

TABLE 4. Ablation study using fastText$_{[MIMIC-III]}$ embedding. For each entry in the table, accuracies of the development and test set are divided by a slash, and the number in parentheses is the best training epoch.

	LSTM	BiLSTM
w/o Attention	0.778 / 0.763 (10)	0.767 / 0.770 (9)[a]
w/ Attention	0.776 / 0.768 (8)	0.779 / **0.778** (12)

[a]This is a group of amended values, and the original values were 0.759 / 0.759 (6).

TABLE 5. More Examples from the MedNLI dataset. P, H, and L stand for premise, hypothesis and label, respectively. Key words for inference are in italics.

Example #4 *Paroxysmal atrial fibrillation.* Patient has abnormal *EKG* **L**: entailment
Example #5 **P**: The patient reports having *reactive airway disease*. **H**: The patient has *COPD* and/or atshma. **L**: entailment
Example #6 **P**: Of note, patient does c/o progressive *shortness of breath* over the past year. **H**: The patient experiences *dyspnea* **L**: entailment
Example #7 **P**: She does admit to some *diarrhea* at home. **H**: The patient is *constipated* **L**: contradiction
Example #8 **P**: Moving all four extremities, *responding* to sternal rub. **H**: Patient is *unresponsive* **L**: contradiction
Example #9 **P**: In the ambulance, per report from the cardiologist, initial rhythm strip showed *normal sinus rhythm*. **H**: The patient had *ventricular fibrillation*. **L**: contradiction
Example #10 **P**: *Liver failure*- hx of encephalopathy, no bx seen in records DM type 2- non insulin dependent CHF Elevated PSA Pancreatitis Postive PPD *Alcoholic cardiomyopathy* **H**: The patient is an *alcoholic*. **L**: neutral
Example #11 **P**: The *PDA* and posterolateral vessels had 90% *ostial lesions*. **H**: Patient may required *CABG* **L**: neutral
Example #11 **P**: A recent *TEE* showed severe *aortic stenosos* with an aortic valve area of 0.7cm2. **H**: Patient has *CHF* **L**: neutral

B. ABLATION STUDY

The main difference between our model and the vanilla ESIM is the newly added *Incorporating Medical Concept Definitions* module: it uses a bidirectional LSTM to encode the definitions of medical concepts, and another attention

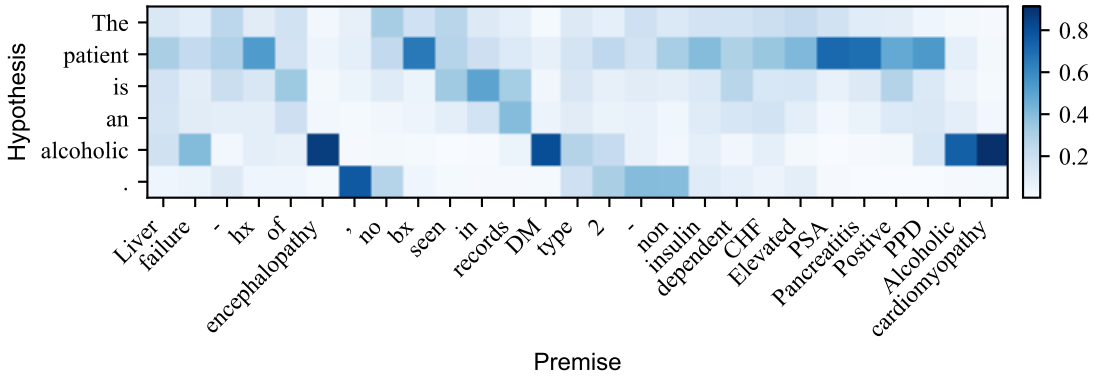

FIGURE 12. Co-attention matrix of Example #10 with normalization over column-axis.

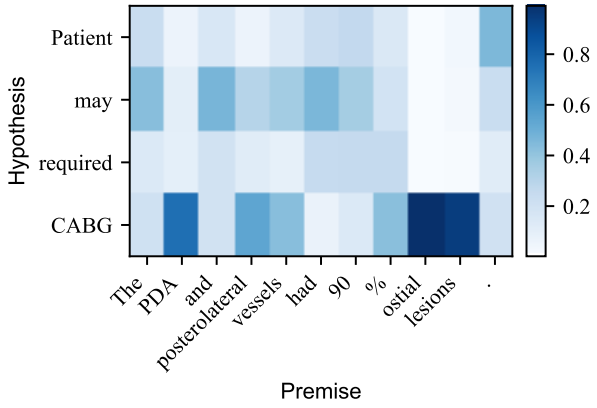

FIGURE 13. Co-attention matrix of Example #11 with normalization over column-axis.

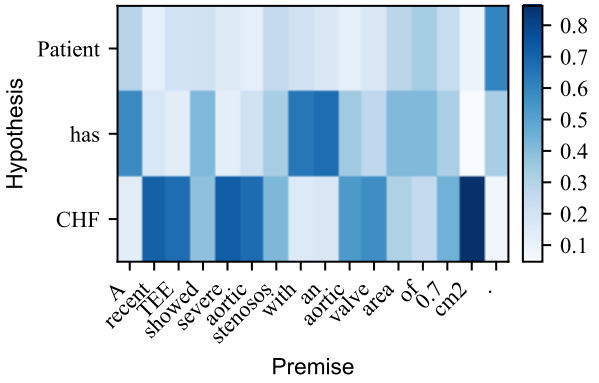

FIGURE 14. Co-attention matrix of Example #12 with normalization over column-axis.

mechanism to enhance vanilla word embeddings. To analyze the contributions of these two components to the overall performance, we conducted an ablation study using fastText$_{[\text{MIMIC-III}]}$ embedding. Three model variants were studied: one that removed only the attention mechanism, another that changed the bidirectional LSTM to unidirectional, and the last that did both. The results of the study are presented in Table 4. The values of model variant w/o attention are amended, because this variant stopped so early compared to others. It was only iterated for 6 epochs, hasn't been fully trained, and did not have good generalization performance in both development set and test set, as shown in Fig. 4. Based on the loss and accuracy curve, we found the 9th epoch was the optimal iteration stop, whose loss was second minimum and best generalization performance.

From Table 4, we can conclude that models w/ attention are better than those w/o attention and bidirectional LSTM is better than unidirectional LSTM. All of these findings reflect the importance of the *Incorporating Medical Concept Definitions* module, and domain-specific knowledge contributes to natural language inference on clinical texts.

C. CASE STUDY

Finally, we qualitatively inspect examples listed in Table 1 and visualize their normalized co-attention matrix, as in (13). For more examples with attention visualizations, see Supplemental material.

The key words of Example #1 for inference are "*diaphoresis*" and "*sweats*". By enhancing word embeddings with knowledge, our model learns to focus on these two medical terms and knows that they have the same meaning. As shown in Fig. 5 (a), in premise, "*diaphoresis*" has the highest weight to "*sweats*". In Fig. 5 (b) (corresponding to Example #2), for the abbreviation "*LP*", our model pays attention to it's full name of "*lumbar puncture*". In Fig. 5 (c) (corresponding to Example #3), our model learns to make inference based on the relationship between "*STEMI*" and "*coronary artery bypass grafting*". Because the definition of "*STEMI*" (i.e., ST segment elevation myocardial infarction) is incorporated, our model learns they are unrelated, and the prediction is neutral class.

VI. CONCLUSION

We have present a novel model for natural language inference on clinical texts by incorporating medical concept definitions into vanilla word embeddings. Our experiment results demonstrated that the model outperforms all baselines, achieving the state-of-the-art performance in accuracy, due to the contributions of domain knowledge.

Further improvement might be made by expanding medical concept definitions dictionary, to cover more medical

terms and abbreviations. For simplicity, we only employed the shortness definition for each concept. However, a concept might have a number of definitions. Therefore, we will study how to encode multiple definitions in the future.

REFERENCES

[1] I. Dagan, O. Glickman, and B. Magnini, "The PASCAL recognising textual entailment challenge," in *Proc. Mach. Learn. Challenges Workshop*. Springer, 2005, pp. 177–190.
[2] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 632–642.
[3] A. Williams, N. Nangia, and S. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2018, pp. 1112–1122.
[4] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 670–680.
[5] Q. Chen, X. Zhu, Z.-H. Ling, S. Wei, H. Jiang, and D. Inkpen, "Enhanced LSTM for natural language inference," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2017, pp. 1657–1668.
[6] Q. Chen, X. Zhu, Z.-H. Ling, D. Inkpen, and S. Wei, "Neural natural language inference models enhanced with external knowledge," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2018, pp. 2406–2417.
[7] A. Romanov and C. Shivade, "Lessons from natural language inference in the clinical domain," in *Proc. 2018 Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 1586–1596.
[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[9] W. Lan and W. Xu, "Neural network models for paraphrase identification, semantic textual similarity, natural language inference, and question answering," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 3890–3902.
[10] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, and C. Zhang, "Disan: Directional self-attention network for RNN/CNN-free language understanding," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 5446–5455.
[11] T. Shen, T. Zhou, G. Long, J. Jiang, S. Wang, and C. Zhang, "Reinforced self-attention network: A hybrid of hard and soft attention for sequence modeling," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Menlo Park, CA, USA: AAAI Press, 2018, pp. 4345–4352.
[12] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a 'Siamese' time delay neural network," in *Proc. 6th Int. Conf. Neural Inf. Process. Syst.*, 1994, pp. 737–744.
[13] S. Wang and J. Jiang, "Learning natural language inference with LSTM," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 1442–1451.
[14] R. Ghaeini *et al.*, "DR-BiLSTM: Dependent reading bidirectional LSTM for natural language inference," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2018, pp. 1460–1469.
[15] D. Bahdanau, K. Cho, and Y. Bengio. (2014). "Neural machine translation by jointly learning to align and translate." [Online]. Available: https://arxiv.org/abs/1409.0473
[16] E. Agirre, A. Gonzalez-Agirre, I. Lopez-Gazpio, M. Maritxalar, G. Rigau, and L. Uria, "SemEval-2016 task 2: Interpretable semantic textual similarity," in *Proc. 10th Int. Workshop Semantic Eval. (SemEval)*, 2016, pp. 512–524.
[17] M. Faruqui, J. Dodge, S. K. Jauhar, C. Dyer, E. Hovy, and N. A. Smith, "Retrofitting word vectors to semantic lexicons," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2015, pp. 1606–1615.
[18] O. Bodenreider, "The unified medical language system (UMLS): Integrating biomedical terminology," *Nucleic Acids Res.*, vol. 32, pp. D267–D270, Jan. 2004.
[19] A. R. Aronson and F.-M. Lang, "An overview of MetaMap: Historical perspective and recent advances," *J. Amer. Med. Inform. Assoc.*, vol. 17, no. 3, pp. 229–236, 2010.
[20] L. Mou *et al.*, "Natural language inference by tree-based convolution and heuristic matching," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2016, pp. 130–136.
[21] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
[22] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Dec. 2017.
[23] G. Tsatsaronis *et al.*, "An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition," *BMC Bioinf.*, vol. 16, no. 1, p. 138, 2015.
[24] A. E. W. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, May 2016, Art. no. 160035.
[25] D. Bahdanau, T. Bosc, S. Jastrzębski, E. Grefenstette, P. Vincent, and Y. Bengio. (2017). "Learning to compute word embeddings on the fly." [Online]. Available: https://arxiv.org/abs/1706.00286
[26] D. Chaudhuri, A. Kristiadi, J. Lehmann, and A. Fischer, "Improving response selection in multi-turn dialogue systems by incorporating domain knowledge," in *Proc. 22nd Conf. Comput. Natural Lang. Learn.*, 2018, pp. 497–507.
[27] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 315–323.
[28] D. P. Kingma and J. Ba. (2014). "Adam: A method for stochastic optimization." [Online]. Available: https://arxiv.org/abs/1412.6980
[29] D. P. Kingma, T. Salimans, and M. Welling, "Variational dropout and the local reparameterization trick," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2575–2583.

MINGMING LU was born in 1991. He received the B.S. degree in computer science from the China University of Mining and Technology, China, in 2013. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology, Tongji University, China. His main research interests include machine learning, natural language processing, and intelligent systems with applications to medicine.

YU FANG received the Ph.D. degree from Tongji University, China, in 2006, where she is currently a Professor with the Department of Computer Science and Technology. Her main research interests include big data analytics and intelligent systems with applications to medicine.

FENGQI YAN was born in 1978. He received the M.S. degree from the Shandong University of Science and Technology, China, in 2007. He is currently pursuing the D.Eng. degree in electronics and information with Tongji University, China. His research interests include medical big data and medical information services.

MAOZHEN LI received the Ph.D. degree from the Institute of Software, Chinese Academy of Sciences, in 1997. He is currently a Professor with the Department of Electronic and Computer Engineering, Brunel University, London, U.K. His main research interests include high performance computing, big data analytics and intelligent systems with applications to smart grid, and smart manufacturing and smart cities. He has over 160 research publications in these areas including four books. He is a Fellow of the British Computer Society and the IET. He has served over 30 IEEE conferences and is on the editorial board of a number of journals.