Integrating Subject-Matter Expertise with Natural Language Inference on Clinical Texts

S.M. Navin Nayer Anik
s.m.navin.nayer.anik@g.bracu.ac.bd

Sabrina Tabassum
sabrinatabassum98@gmail.com

Fahim Abrar
fahim.abrar1@g.bracu.ac.bd

Md Farhadul Islam
md.farhadul.islam@g.bracu.ac.bd

Md Sabbir Hossain
md.sabbir.hossain1@g.bracu.ac.bd

Annajiat Alim Rasel
annajiat@gmail.com

***Abstract*—The approach that makes use of clinical literature to arrive at findings is one that has not been exposed to a substantial amount of examination. This effort, which is extremely appreciative for it, is receiving a boost from the recently published and painstakingly annotated MedNLI dataset. The clinical writing language is distinguished from other types of written language by a number of distinct characteristics. These characteristics include an abundance of medical terminology and acronyms, as well as a diversity of written forms for the same medical notion. When all of these factors are combined, it is far more difficult to make conclusions from clinical literature than it is from data that is readily accessible to the public. In this paper, we assemble a new incorporating medical concept denitions 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 denition representations, and finally attends these denition representations. This process is repeated until the most relevant medical concept is found for each word. The improved sequential inference model, often known as ESIM, serves as the foundation for this module. Incorporating information that is specific to the situation at hand is one method that may be used to successfully eliminate this problem. The purpose of the empirical assessments included in this research is to demonstrate that our model is able to improve prediction performance and achieve a high level of accuracy when it is applied to the MedNLI dataset. This will be accomplished by providing evidence that our model can achieve these two goals. To elaborate a little bit further, the knowledge-enhanced word representations provide the entailment class a sizeable performance increase.**

***Index Terms*—Attention mechanism, clinical text, medical domain knowledge, natural language inference, word representation.**

I. Introduction

A job that addresses the semantic link (entailment, contradiction, or neutral) between a premise and a hypothesis is referred to as Natural Language Inference (NLI), which is also known as Recognizing Textual Entailment (RTE). The term ”Recognizing Textual Inference” (RTI) is sometimes used interchangeably with ”Natural Language Inference.” The availability of large-scale annotated datasets to the general public has increased over the last several years, which has accelerated up the growth of this activity. These datasets are represented by two different corpora: the Stanford Natural Language Inference (SNLI) [2] corpus and the Multi-Genre Natural Language Inference (MultiNLI) [3] corpus. In addition, quite a few different models of deep neural networks have been offered in order to reach state-of-the-art performance [4]. These models have been developed in order to fulfill certain goals. These models have been built in order to execute different jobs.[6]. The clinical domain is significant to the recently published MedNLI [7] dataset, the primary emphasis of which is on NLI tasks carried out on clinical texts. Because of the novelty and specificity of this field, the language phenomena that are found in clinical texts are distinct from those that are observed in data from the open domain. This is because clinical texts are written for a very specialized audience. Some instances of these are included below: (1) The presence of a high number of medical terminology and abbreviations contributes to the problem of out-of-vocabulary (OOV); (2) A same medical notion may be articulated in a variety of different ways across vocabularies, even though they all transmit the same meaning. The MedNLI dataset is the source for some of the examples included in Table 1. These illustrations are being supplied for illustrative purposes only. Both ”sweats” and ”diaphoresis” relate to the same process that occurs in a person’s body, despite the fact that their names are spelled differently. The process of sweating is referred to as diaphoresis in the medical field. These two ideas are required for achieving a sufficient level of knowledge with regard to the first example. The second and third examples both include common medical acronyms (”LP” and ”STEMI”), as well as common medical terminology (”coronary artery bypass grafting” and ”lumbar puncture”), but they do not contain any typical logogram words (”pt”, which stands for patient). In the event that a computer system is unable to interpret certain medical terminology and acronyms in the proper manner, it is possible that the system would incorrectly classify patients. In general, the availability of these one-of-a-kind language occurrences makes it a great lot more tough to arrive at conclusive conclusions on MedNLI. This is because MedNLI is a corpus that contains a significant number of information. In this study,

we integrate domain-specific information into a standard open domain model (ESIM) by encoding the definitions of medical concepts using a bidirectional long short-term memory [8] (BiLSTM), and then attending the vanilla word embeddings to these domain-specific representations. Specifically, we achieve this by encoding the definitions of medical concepts using a bidirectional long short-term memory [8] (BiLSTM). This is accomplished by the use of a bidirectional long short-term memory [8] (BiLSTM) for the purpose of encoding the meanings of medical ideas. To be able to process clinical texts, one needs knowledge that is specific to the clinical sector; as a result, possessing this information is necessary. It is possible to instruct computers in this manner so that they can, on the one hand, grasp the meanings of medical terminology and acronyms and, on the other hand, discover similarities and differences between a range of medical ideas. Experiments are carried out using the MedNLI dataset, and the results demonstrate that our model achieves a level of performance that is considered to be state-of-the-art. It also surpasses all baselines provided by Romanov and Shivade [7]. In addition, we provide you both an ablation study and a case study so that you can have a better grasp of the role that domain knowledge plays in the process of the creation of our model. We do this so that you can have a more complete picture of how our model is constructed. The following are the three most significant contributions that may be taken away from our work: We make accessible a knowledge-augmented model for natural language inference on clinical texts. This model is used to draw conclusions from clinical material. This model incorporates both the BiLSTM algorithm and attention in order to enhance standard word embeddings with the meanings of medical subject areas. Inferences drawn from natural language analysis of clinical literature served as the impetus for the creation of this methodology. When we test the utility of our model on the MedNLI dataset, we find that it is able to achieve a greater degree of accuracy than models that do not have their data supplemented with knowledge. This is something that we discovered when we tested the usefulness of our model. The results of our research on ablation, as well as the findings of the case study that we carried out, give some novel and intriguing insights into the contributions of knowledge-enhanced word representations. The remaining aspects of this work have been structured in accordance with the description that is provided below. In the next part (part II), we will discuss the amount of work that is required to make inferences based on natural language. In the third section, the proposed model is taken apart into its component parts so that each one may be investigated in more depth. In Section IV, we go into great depth on the experimental set-up, and then in Section V, we give a review of both the experimental circumstances and the outcomes. In the next section (Section VI), the conclusion is discussed. Our results are reported in the last section, which is designated as part VI of this document.

II. Related Works

Encoding-based models and interaction-based models are the two basic kinds of techniques that have been developed for the aim of conducting natural language inference [9]. Encoding-based models focus on how words are encoded, while interaction-based models examine how words are used in context. Encoding-based models concentrate on the process by which words are encoded, while interaction-based models investigate the ways in which words are used in different settings. Encoding-based models focus on the process by which words are encoded, while interaction-based models study the manner in which words are employed in a variety of contexts. Encoding-based models are more common. Employing encoding-based models [2, 4], [10], and [11] in conjunction with siamese architecture [12] makes it possible to acquire the information necessary to learn vector representations of the premise and the hypothesis. This may be done in a viable manner. One will be able to get the necessary knowledge as a result of this. When they have completed this stage of the process, the next thing that they do is make use of a neural network classier in order to determine whether or not there is a semantic relationship between the two phrases. InferSent [4, which is one of the baseline models that are included in the MedNLI dataset, is an example of the kind of model that would be considered to be representative of the norm. This model is a great depiction of what the typical situation may look like. The following are some more models that are representative in their own right and serve as examples for those models. After using a number of different tactics for word alignment, such as attention [15], models that are based on interactions [5, 13, 14] are used to aggregate inter-sentence interactions. [5, 13], and [14] are the best places to look for the references that are associated with these models. According to the results of the SemEval-2016 challenge of interpretable semantic textual similarity [16], it would seem that the semantic relations of aligned chunks give a significant contribution to the modeling of phrase pairs. This conclusion may be drawn from the outcomes of the challenge. The results of the competition including interpretable semantic textual similarity served as the basis for these conclusions. In addition, models that are based on interaction perform notably better than those that are based on encoding when it comes to the correctness of the outcomes that they provide. Encoding-based models are much less accurate. Enhanced Sequential Inference Model (ESIM) is the name that Chen and his colleagues [5] have given to their proposal for a new model that they have built. [5] ESIM stands for ”enhanced sequential inference model.” [5] Chen and his associates are responsible for the development of this concept. The input encoding, the matching of co-attention, and the inference composition are the three main components that make up this system as a whole. In addition to the other baseline models that are included in the dataset, the ESIM baseline model is one of the baseline models

that is included in the MedNLI dataset. In contrast to previous research [6], which enriched NLI models with lexical-level semantic information about synonymy, antonymy, hypernymy, hyponymy, and co-hyponymy between words, the research that is presented here focuses on the medical domain and investigates the incorporation of additional knowledge on clinical texts for natural language inference. Specifically, this research looks at the incorporation of additional knowledge on clinical texts for natural language inference. In particular, the purpose of this study is to investigate the inclusion of new information on clinical texts for the sake of natural language inference. In specifically, the goal of this research is to evaluate the potential benefits that may be gained by natural language inference from the addition of additional information. More explicitly, the purpose of this study is to assess the possible advantages that may be achieved by natural language inference from the inclusion of extra information. explicitly, the objective of this research is to evaluate the potential benefits that may be gained by natural language inference. In addition to this, Romanov and Shivade [7] investigated two other ways that may be used in order to include domain-specific data into their baseline models. Both of these strategies have the potential to be utilised. In one sense, in order to make it possible for the input to models to represent clinical information, they updated the pre-trained word embeddings using a technique known as retroting [17]. This was done in order to make it feasible for the input to models to convey clinical information. This was done in order to ensure that the clinical information could be properly sent by the input to the models. This was done so that the information could be successfully transferred from the input to the models. This was done in order to make the successful transfer of information feasible. This operation was carried out in order to meet the goals that had been determined before the process began. Despite this, continuing in this manner will only aid to bring about a decrease in performance if it is continued for a longer period of time. The decline in performance is going to be brought about if this conduct is allowed to continue. It is required to engage in additional cognitive processes in order to arrive at conclusions that are appropriate for use in the medical field. This is due to the fact that the retroting process can only be used to concepts that are relatively close to one another, but medical ideas are often more complicated. One such tactic is to direct one's emphasis toward the accumulation of one's own body of knowledge. Because it offers a contribution that is beneficial to both the InferSent model and the ESIM model, this has a number of favorable implications for both of these models. One method for doing this is to focus one's attention on the knowledge that one already has and has access to. However, in contrast to the first technique, which involves making modifications to the model's inputs, our model takes use of definition representations in order to improve the word embeddings of medical abbreviations and phrases. This is done in order to reduce the number of false positives and increase the number of correct predictions. Because of this, our system has a greater degree of accuracy when it comes to predicting the meaning of medical terms. Because of this, we have a better understanding of the medical facts, which is one of the contributing factors. This helps to lessen the OOV problem and bridges the semantic gap that now exists between the many textual representations of a medical thought that are now accessible in this day and age. This is because multiple textual representations of a medical notion are now available in this day and age.

III. Methodology

A. *DOMAIN KNOWLEDGE*

In order to retrieve the definitions of the various medical terms, we first look them up in the Unified Medical Language System (UMLS), which is located at [18]. Each medical concept that was included in the UMLS would be accompanied by a number of definitions, with each group of definitions being derived from a different source vocabulary. We decided to go with the most concise explanation of this medical principle so that the model would be easier to comprehend. This allowed us to more effectively communicate its purpose. Our K knowledge base, which is an abbreviation for domain knowledge base, contains a grand total of 198,042 definitions once everything is said and done.

Second, in accordance with the work that has been done in the past, we utilise Metamap to extract medical ideas from premise and hypothesis statements. After that, we map those concepts to standard terminologies that are contained in the UMLS. It is possible for each extracted phrase to be linked to more than one idea, and the concepts that are linked to each phrase are organised in descending order according to the score they received on the MetaMap Indexing (MMI) test. If the score is higher, it indicates that the medical concept being discussed has a greater degree of connection to the phrase from which it was taken. In the course of this investigation, we are only going to take into consideration the concept that was awarded the highest score for each category, and we are going to disregard any others that were awarded a lower score. As a direct result of this, each and every word is associated with either zero or one medical concept. Now that we have a comprehensive comprehension of the idea that the medical term or abbreviation refers to, we are able to map a range of textual forms to the same concept. The very last step, but certainly not the least important, is to link the concepts to the words that best describe them. For example, if a single word (ai) in the premise phrase can be used to extract a medical concept, we will look for the definition of that concept in our knowledge base (K) that is specific to the industry in which we operate.

B. MODEL OVERVIEW

In this section, we will discuss the methodology that we have developed for making inferences from clinical papers using natural language. Its structure is made up of several layers, including the input encoding layer, the co-attention layer, and the inference composition layer. The model receives as inputs the sentence that functions as the premise, the sentence that functions as the hypothesis, as well as the definitions of the medical terms that were obtained from two separate statements. The model will then begin by generating relevant word representations by utilising word embeddings that have already been trained. These pre-trained word embeddings may be word embeddings that are openly accessible to the public and are in the public domain, or they may be word embeddings that have been trained on a corpus that is specific to a certain area. After that, the module titled "Incorporating Medical Concept Definitions" will look at each word in the two sets of phrases in great detail to determine whether or not they have a definition. The module will make use of an existing definition if one is available. In addition, by feeding the enhanced word embeddings into a Siamese BiLSTM network, one can generate a collection of contextualised representations of premise and hypothesis sentences.

Following the completion of a heuristic matching technique in order to collect local inference vectors for each word, we first perform a soft-alignment of contextualised word representations between the premise and the hypothesis in order to obtain an aligned representation in the co-attention matching layer. This is done in order to obtain an aligned representation in the co-attention matching layer. In the very last stage of the procedure, another BiLSTM is employed to combine the gathered local inference vectors as a component of the inference composition layer. This step is the culmination of the process. This phase is required in order to determine the overall inference connection that exists between the premise and the hypothesis. The relationship might be either positive or negative. The output hidden vectors of the second BiLSTM are first max pooled, then mean pooled, and lastly transformed to fixed-length vectors in order to determine the inference class. This process is carried out in order to establish the inference class. After that, these vectors are sent into the last iteration of the multilayer perceptron (MLP) classifier.

C. INPUT ENCODING LAYER

The premise, the hypothesis, and the corresponding medical concept definitions are all accepted as inputs by the input encoding layer, where may be replaced with i or j. The input encoding layer also takes all of these things into consideration when determining the output. Initially, pre-trained word embeddings are applied to the word inputs in order to convert them into vector sequences. After that, the breadth and depth of the lexicon are taken into account. it is a component of the dimension that the word embedding has. Throughout the course of the trials, we investigate a total of six distinct word embeddings. This includes one that is openly licenced and available to the general public, two that have been trained using domain-specific corpora, and three that have been initialised using open-licensed word embeddings and then further fine-tuned using one or two domain-specific corpora.

GloVe[CC] is an abbreviation for the GloVe embeddings that were taught using Common Crawl. • fastText[BioASQ]: embeddings created using fastText , which were trained using PubMed abstracts from the BioASQ challenge. • fastText[MIMIC-III]: embeddings of fastText that have been trained using patient clinical notes taken from the MIMIC-III database . • GloVe[CC] fastText[BioASQ]: GloVe embeddings for initialization, which may then be further fine-tuned based on the BioASQ data. • GloVe[CC] fastText[BioASQ] fastText[MIMIC-III]: GloVe embeddings for initialization and further finetuning on the BioASQ and MIMIC-III data in succession. GloVe embeddings for initialization and further fine-tuning on the BioASQ and MIMIC-III data. • fastText[Wiki], fastText[MIMIC-III]: fastText Wikipedia embeddings for initialization and subsequent fine-tuning based on the MIMIC-III data.

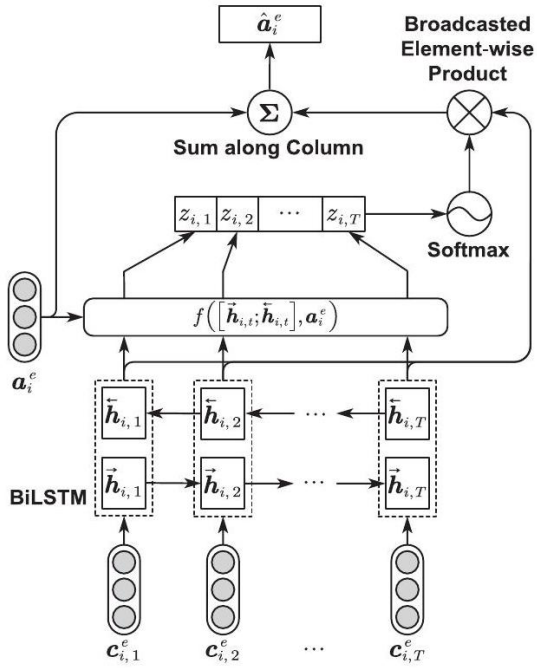
D. INCORPORATING MEDICAL CONCEPT DEFINITIONS

We are able to provide definitions of medical ideas by embedding them inside words as a direct consequence of being motivated by the work of. This is seen in the. The bidirectional long short-term memory network, also known as the BiLSTM network, has been proven to be successful at representing dependencies in sequences that arise from both the past and the future. This has been shown by the fact that the BiLSTM network is commonly referred to. As a direct consequence of this, we make use of it in both the forward and the backward direction to encode definition embeddings. The input that is given to the BiLSTM at the time step t is represented by the letter c. Because including it here would make the notation more difficult to understand, the subscript i has been omitted from this section. In the forward motion, the following are the changes that are made to the hidden states:

$$
\begin{aligned}
\boldsymbol{i}_t &= \sigma(\boldsymbol{W}^i \boldsymbol{c}_t^e + \boldsymbol{U}^i \overrightarrow{\boldsymbol{h}}_{t-1} + \boldsymbol{b}^i) \\
\boldsymbol{f}_t &= \sigma(\boldsymbol{W}^f \boldsymbol{c}_t^e + \boldsymbol{U}^f \overrightarrow{\boldsymbol{h}}_{t-1} + \boldsymbol{b}^f) \\
\boldsymbol{o}_t &= \sigma(\boldsymbol{W}^o \boldsymbol{c}_t^e + \boldsymbol{U}^o \overrightarrow{\boldsymbol{h}}_{t-1} + \boldsymbol{b}^o) \\
\boldsymbol{q}_t &= \tanh(\boldsymbol{W}^q \boldsymbol{c}_t^e + \boldsymbol{U}^q \overrightarrow{\boldsymbol{h}}_{t-1} + \boldsymbol{b}^q) \\
\boldsymbol{p}_t &= \boldsymbol{f}_{i,t} \circ \boldsymbol{p}_{i,t-1} + \boldsymbol{i}_t \circ \boldsymbol{q}_t \\
\overrightarrow{\boldsymbol{h}}_t &= \boldsymbol{o}_t \circ \tanh(\boldsymbol{p}_t)
\end{aligned}
$$

The hidden state at time step t is formed not only by the input word, but also by the hidden state that happened immediately before it. This means that the input word alone is not sufficient to determine the hidden state. In the same

way, while going in the other direction, the hidden state h is changed by taking the most recent input into account in addition to the hidden state from the next time step. This occurs when traveling backwards in time. This takes place while the vehicle is going in reverse. The output of the BiLSTM is typically acquired at the t-th time step by concatenating the hidden states from both directions. This is done in a typical implementation. This takes place during the phase known as the bidirectional concatenation. The t-th time step in the process sees this particular event take place. To be more explicit, the process that was given may be simplified into the BiLSTM function, which can be written as follows:



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

In order to construct definition-enhanced word embeddings, we first aggregate the outputs of the BiLSTM by using a method known as multi-layer perceptron attention [15], and then we add those outputs to the word embeddings that were produced by the vanilla approach. This process is repeated until we have successfully constructed the definition-enhanced word embeddings. Because of this, we are in a position to acquire word embeddings with enhanced definitions. In order to get a more accurate result, attention performs the computations listed below in order to determine the alignment score:

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

In the following words, a weight vector shall be referred to as v whenever weight matrices are used; this notation will be used throughout.This alignment score gives an indication of how attentive you are to the activity that is currently being performed. After that, a softmax function that standardizes the scores is employed in order to generate an alignment score vector. This is done in order to meet the requirements of the alignment. This method is invoked after the prior stage in the process has been finished successfully.

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

A shared-parameter BiLSTM is fed the enhanced premise and hypothesis word embeddings to produce context-specific representations of words.

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

E. CO-ATTENTION MATCHING LAYER

The stage at which a model of the interactions is produced is the step that is the most important when it comes to determining how the premise and the hypothesis are related to one another in terms of inference. This is because the stage at which a model of the interactions is constructed is where the most important information is found. The dot-product operation is used to construct aligned word representations in this layer, which ultimately leads to the formation of a co-attention matrix. After that, matching information is acquired at the word level by comparing the contextualized representations of the aligned words with the representations of the contextualized words. This is done so that the matching information may be retrieved at the word level. To get things rolling, the first thing we do is use the following formula to get the co-attention score between each representation tuple.

$$e_{ij} = (\boldsymbol{a}_i^s)^T \boldsymbol{b}_j^s$$

F. INFERENCE COMPOSITION LAYER

Within this layer, the aggregation approach that is often applied to create the local inference vectors that were gathered above is a parameters shared BiLSTM that is followed by maximum and mean pooling operations:

Once again, we make use of the BiLSTM, but this time its function is quite different from what it was previously used for. In this example, the BiLSTM is educated to differentiate between important local inference vectors in order to acquire the overall sentence-level inference connection that exists between the premise and the hypothesis. This is done in order to

$$
\begin{aligned}
\boldsymbol{a}_1^v, \ldots, \boldsymbol{a}_M^v &= \text{BiLSTM}_2(\boldsymbol{a}_1^m, \ldots, \boldsymbol{a}_M^m) \\
\boldsymbol{b}_1^v, \ldots, \boldsymbol{b}_N^v &= \text{BiLSTM}_2(\boldsymbol{a}_1^m, \ldots, \boldsymbol{b}_N^m) \\
\boldsymbol{a}_{max}^v &= \max_{1 \leqslant i \leqslant M} \boldsymbol{a}_i^v \\
\boldsymbol{a}_{mean}^v &= \text{mean}_{1 \leqslant i \leqslant M} \boldsymbol{a}_i^v \\
\boldsymbol{b}_{max}^v &= \max_{1 \leqslant j \leqslant N} \boldsymbol{b}_j^v \\
\boldsymbol{b}_{mean}^v &= \text{mean}_{1 \leqslant j \leqslant N} \boldsymbol{b}_j^v
\end{aligned}
$$

learn the relationship between the premise and the hypothesis. This is done so that the connection between the premise and the hypothesis may be better understood. The pooling vectors, after being concatenated, are then input into the final iteration of the multi-layer perceptron (MLP) classifier. This classifier consists of one hidden layer, tanh activation, and a softmax output layer:

$$
\boldsymbol{y} = \text{MLP}([\boldsymbol{a}_{max}^v; \boldsymbol{a}_{mean}^v; \boldsymbol{b}_{max}^v; \boldsymbol{b}_{mean}^v])
$$

G. OPTIMIZATION OBJECTIVE

To obtain the maximum possible decrease in the total amount of multi-class cross-entropy loss is the goal of the training process, which spans the whole of the model. This should be as close to zero as is practically possible. The following is a description of what is meant when we talk about ”loss function”:

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

where represents all trainable parameters, D is the number of training instances, and yi is the ground truth for the i-th example in the training set. where D is the number of training examples. where yi represents the fundamental reality of the ith illustration in the training set.

IV. EXPERIMENTS

A. MEDNLI DATASET

We assessed our model using the MedNLI dataset [7], which consists of 13 thousand phrase pairs that have been expertly annotated. The clinical notes that are included in the MIMIC-III v1.3 database [24] were used to compile the premise sentences, while four different doctors came up with the statements that make up the hypothesis. The end outcome is that

The dataset contains a total of 14,049 different premise and hypothesis combinations. There are 11,232 pairs that are designated for training, 1,395 pairs that are designated for development, and 1,422 pairs that are designated for testing. The average sentence length of a premise is 20 words, whereas

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

the average sentence length of a hypothesis is 5.8 words. In the meanwhile, the maximum number of words allowed for premises is 202, and the maximum number of words allowed for hypotheses is 20. We utilize the same data split that is given in Romanov and Shivade [7], and classification accuracy is the parameter that we use to evaluate the system.

B. TRAINING DETAILS

We opted to use a dimension of 300 for the word embeddings and hidden states of BiLSTMs in accordance with the configurations of all of the baselines on the MedNLI dataset. With one exception, the BiLSTM that was used in the module for integrating medical concept definitions had a value of 150. This was because the BiLSTM was utilized in the context of the incorporation of medical concept definitions. We restricted the length of the phrases that comprised the premise and the hypothesis to a maximum of 50 words, but the length of the sentences that explained medical terms was restricted to a maximum of 200 words. Throughout the course of training, every single word embedding was secured in its position. Adam [28] was used for the process of optimizing the system, and its starting learning rate was set at 0.001. The size of the mini-batch will be 64, since it was the amount that was decided upon. We came to the conclusion that a dropout rate of 0.5 should be applied to both the input and output of the hidden layer in the final version of the MLP classifier. In addition to that, we employed variational dropout [29] as the input to the BiLSTMs, and the value for this parameter was also set to 0.5. We let our model be trained for a total of twenty epochs while it was in our care. When, after five subsequent epochs, the development loss did not show any symptoms of improving, the training was halted since it was deemed futile. The test set that corresponds to the development set was then used to put each hyper-parameter through its paces after it had been meticulously selected on the development set. PyTorch2 and AllenNLP3 were the primary tools that we used in the development of our model.

V. RESULT AND DISCUSSION:

COMPARISON AGAINST BASELINES

We test our model, which is known as ESIM w/ Knowledge, against the InferSent and ESIM baseline models that were evaluated by Romanov and Shivade [7] for the six distinct word embeddings that are described in Section III-D. Table 3 presents the findings in their entirety. Our model surpasses all baseline models and

achieves the performance of the state-of-the-art system, demonstrating that the incorporation of medical concept defi-

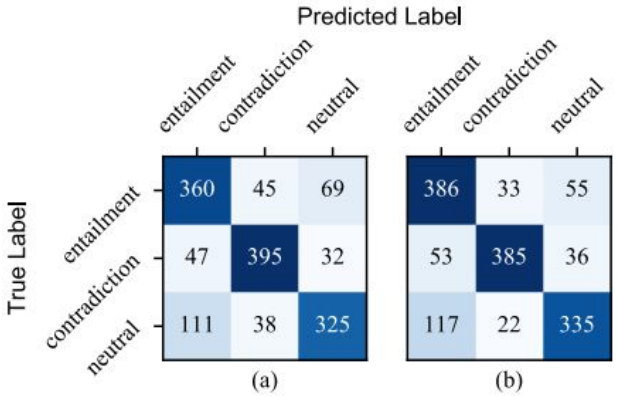
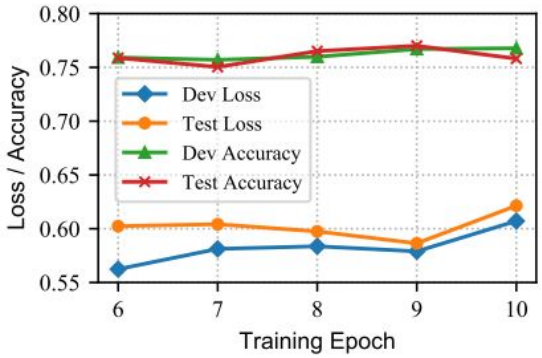


Fig. 1.



nitions may greatly enhance the system's overall performance. We found an absolute gain of 0.012, which corresponds to a 1.6% relative increase in the model utilizing fastText[MIMIC-III] embedding when compared to the best baseline (i.e., InferSent using fastText[Wiki] fastText[MIMIC-III] embedding). This was in comparison to the InferSent model, which used fastText[Wiki] embedding. In point of fact, a total of three outcomes for various word embeddings (the others being fastText[Wiki] fastText[MIMIC-III] embedding and GloVe[CC] fastText[BioASQ] fastText[MIMIC-III] embedding) surpass the best baseline.

In baseline models, all of InferSent's findings are superior than those obtained by ESIM, with the exception of one. However, when compared to either of the two baselines, our results for each word embedding are superior, which demonstrates the efficiency of ESIM when combined with domain knowledge. Our model achieves its maximum success with the GloVe[CC] fastText[BioASQ] embedding (0.765 as opposed to 0.745), where we receive an absolute gain of 0.02 and a relative gain of 2.7%. [CC] stands for the GloVe Coordinate Conversion.

In addition to analyzing the performance as a whole, we also build the confusion matrix to display the classification results of the three different classes (neutral, entailment, and

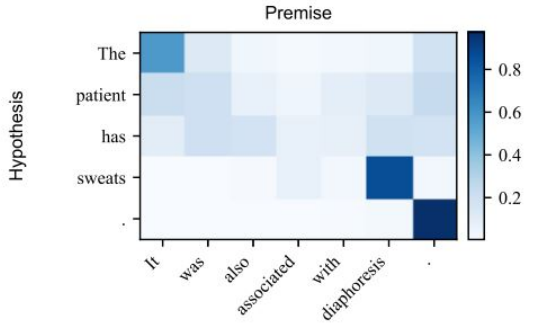


Fig. 2.

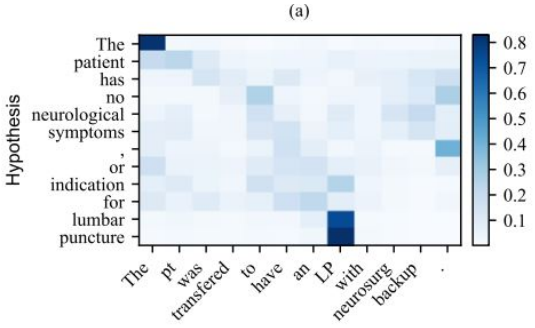


Fig. 3.

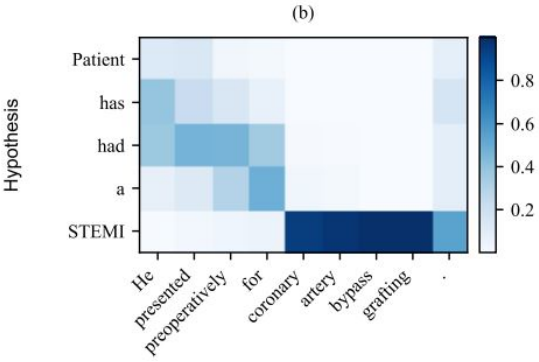


Fig. 4.

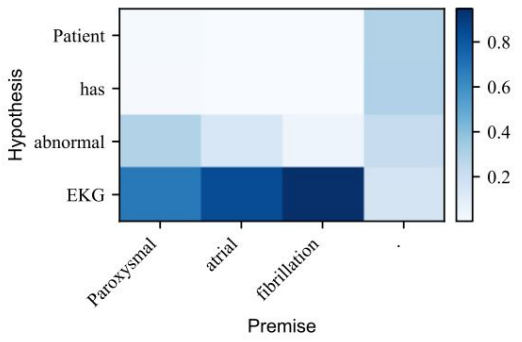


Fig. 5.

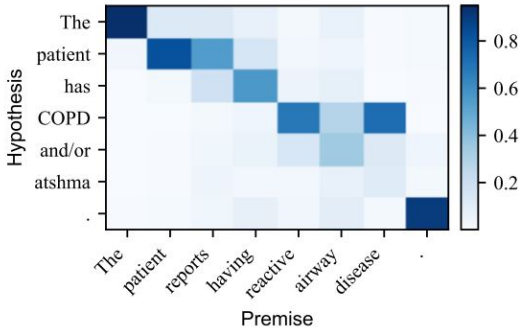


Fig. 6.

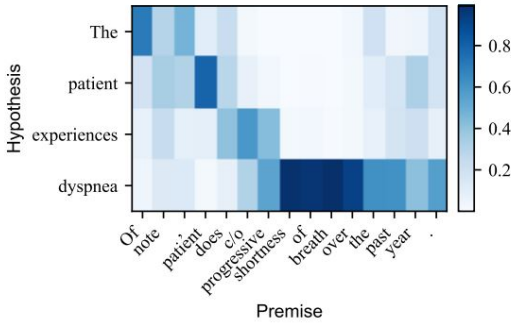


Fig. 7.

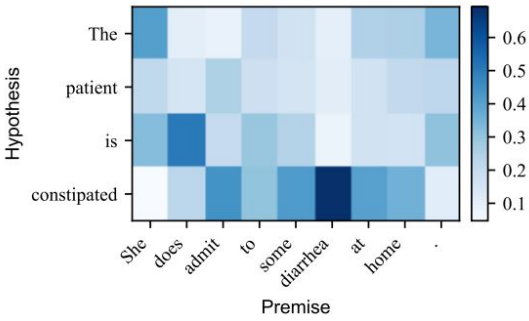


Fig. 8.

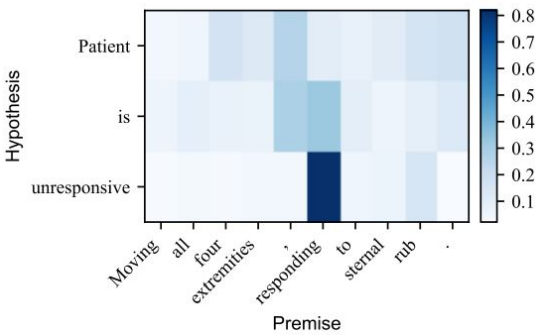


Fig. 9.

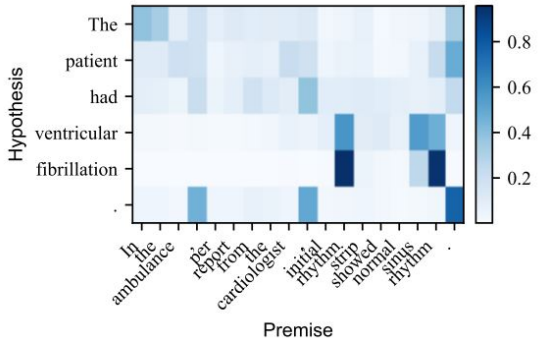


Fig. 10.

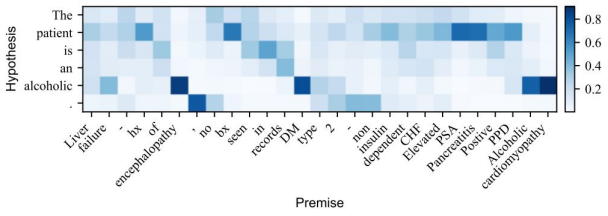


Fig. 11.

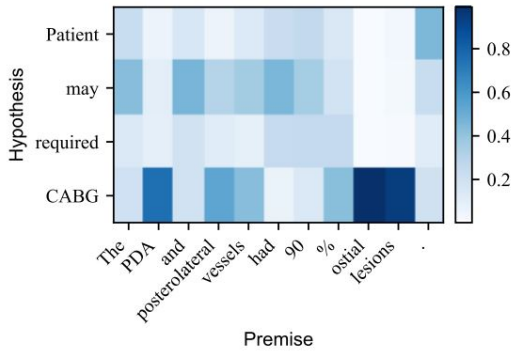


Fig. 12.

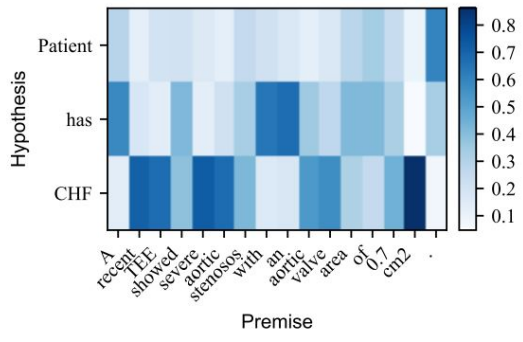


Fig. 13.

contradiction). As can be seen in Figure 3, there are two confusion matrices that have not been normalized. The one on the left is associated with the best baseline4, and the one on the right is associated with the best outcome produced by our model. When these two confusion matrices are compared to one another, one might arrive at the following conclusions:

1. Our approach increases the performance in the entailment and neutral classes, with a significant contribution made to the entailment class. Additionally, the misclassifications made to the contradiction and neural classes are decreased by 12 and 14, respectively. Our hypothesis is that this is the case because the included domain knowledge improves the word representations of medical words and abbreviations and bridges the semantic gap between the various written versions of the same medical idea. A further benefit of incorporating this information is that it lowers the likelihood of the neural class being incorrectly categorized as the contradiction class.

2. When it comes to the battle for the contradiction class, our model comes out on top. After reviewing the examples that resulted in incorrect classification, we found that the ones that required the use of numerical reasoning were the ones in which errors occurred the most often. For example, the statement "In the ED, initial VS revealed T 98.9, HR 73, BP 121/90, RR 15, O2 sat 98% on RA" is incorrect because it includes a mistake. Our model often confuses situations involving numerical reasoning with entailment class because of its predisposition to do so. In the neuronal class, the same may be said about the situation. We think that by using ensemble methodologies like InferSent and ESIM with Knowledge, we will be able to take advantage of the characteristics that each model has, which will ultimately lead to higher prediction performance.

VI. Conclusion

A novel model for natural language inference based on clinical texts has been provided by our team. This model was created by incorporating definitions of medical concepts into simple vanilla word embeddings. We were the ones that came up with this model. The results of our studies shown that the model is superior to all baselines and reaches the state-of-the-art performance in terms of its accuracy. This was made possible as a direct consequence of the contributions of domain knowledge. It's possible that more progress may be made by expanding the scope of the dictionary of medical concept definitions to cover a bigger number of medical terms and abbreviations. This would be a step in the right direction. Because we want for everything to be as simple as possible, we limited ourselves to using just the most succinct definitions for each topic. Having said that, depending on the circumstances, a particular concept may be understood in more than one way. As a result of this, in the near future we are going to look at other methods that multiple meanings may be encoded.

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] 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.

[5] 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.

[6] 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.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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

[15] 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.

[16] 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.

[17] O. Bodenreider, "The unified medical language system (UMLS): Integrating biomedical terminology," Nucleic Acids Res., vol. 32, pp. D267–D270, Jan. 2004.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.
[23] A. E. W. Johnson et al., “MIMIC-III, a freely accessible critical care database,” Sci. Data, vol. 3, May 2016, Art. no. 160035.
[24] X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier neural networks,” in Proc. 14th Int. Conf. Artif. Intell. Statist., 2011, pp. 315–323.
[25] 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.