Improving Medical NLI Using Context-Aware Domain Knowledge

Shaika Chowdhury
Department of Computer Science
University of Illinois at Chicago
schowd21@uic.edu

Philip S. Yu
Department of Computer Science
University of Illinois at Chicago
psyu@uic.edu

Yuan Luo
Department of Preventive Medicine
Northwestern University
yuan.luo@northwestern.edu

Abstract

Domain knowledge is important to understand both the lexical and relational associations of words in natural language text, especially for domain-specific tasks like Natural Language Inference (NLI) in the medical domain, where due to the lack of a large annotated dataset such knowledge cannot be implicitly learned during training. However, because of the linguistic idiosyncrasies of clinical texts (e.g., shorthand jargon), solely relying on domain knowledge from an external knowledge base (e.g., UMLS) can lead to wrong inference predictions as it disregards contextual information and, hence, does not return the most relevant mapping. To remedy this, we devise a **knowledge adaptive** approach for medical NLI that encodes the premise/hypothesis texts by leveraging supplementary external knowledge, alongside the UMLS, based on the word contexts. By incorporating *refined* domain knowledge at both the lexical and relational levels through a **multi-source attention mechanism**, it is able to align the token-level interactions between the premise and hypothesis more effectively. Comprehensive experiments and case study on the recently released MedNLI dataset are conducted to validate the effectiveness of the proposed approach.

1 Introduction

Natural Language Inference is a fundamentally important but challenging task in Natural Language Processing (NLP) as it requires understanding and reasoning over natural language texts (MacCartney and Manning, 2009). As a result, a good performing NLI system is considered indispensable for downstream NLP applications such as question answering and automatic text summarization (Harabagiu and Hickl, 2006; Lloret et al., 2008). Given a pair of sentences, a premise p and a hypothesis h, the goal of NLI is to determine whether the semantic relationship between p and h is among $entailment$, $contradiction$ and $neutral$.

The ability to understand natural language text innately requires to deal with background knowledge [1] (Long et al., 2017; Weissenborn et al., 2017). A robust NLI model usually needs to reason over two types of background knowledge - lexical and relational (Weissenborn et al., 2017). The former pertains to understanding the concepts expressed by the words in the text, while the latter learns the semantic relations between the different concepts. When performing NLI on open domain data, it is assumed that the background knowledge will be implicitly learned from the training corpora. Re-

Premise: - **DMII** complicated by **DM** neuropathy - PVD s/p L CFA w/balloon angioplasty of SFA and AK [**Doctor Last Name **] artery w/ persistent non-healing **ulcer** at the lateral and medial malleolus, non-healing L pedalulcer - **Hypertension** - h/o MDR Pseudomonas and MRSA **skin infections** - h/o hemorrhagic pancreatitis ([**2857**]) - h/o cholecystitis (still has gallbladder)

Hypothesis: Patient has multiple **diabetes** related **comorbidities**.

Label: Entailment

Figure 1: Sample premise-hypothesis pair from MedNLI. The words in red, “DMII" and “DM" in p and “diabetes" in h are semantically similar at the lexical level. The UMLS relation “co-occurs" of the highlighted words in green in p to “diabetes" in h manifests the inferential signal “comorbidities".

[1] background/external/domain knowledge are used interchangeably in this paper

*Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 1–11
Barcelona, Spain (Online), December 12–13, 2020

lease of large NLI datasets like the Stanford Natural Language Inference (SNLI) (Bowman et al., 2015) and Multi-Genre Natural Language Inference (MultiNLI) (Williams et al., 2017) corpora, with around 570,000 and 433,000 sentence pairs respectively, have made it possible to train deep neural networks, which are capable of encoding this knowledge in their parameters.

For specialized domains (e.g., medical), however, large NLI datasets are extremely scarce and the required implicit knowledge beyond the text surface cannot be extracted from limited data. For example, the recently released MedNLI dataset (Romanov and Shivade, 2018), albeit being the single publicly available NLI dataset in the clinical domain, contains only around 13,000 expert annotated sentence pairs [2]. Therefore, most current literature on medical NLI (Romanov and Shivade, 2018; Jin et al., 2019) have capitalized on the prior semantic knowledge that is encoded in external resources (e.g., UMLS). Nevertheless, a limitation of the existing suite of medical ontologies such as UMLS is that they retrieve mapping (i.e., lexical/relational) irrespective of the textual context, which could mislead the inference model. This is worsened by the distinct linguistic idiosyncrasies present in clinical texts, wherein phrases are compressed with shorthand jargon (i.e., abbreviations) for physicians' convenience. Specifically, this instigates three challenges: 1) some semantically important words do not map to any matching concept in the UMLS [3], 2) a wrong concept mapping is returned that does not reflect the word's actual meaning and 3) the noise introduced by wrong concept mapping could get carried forward when retrieving the relational mapping.

To address the aforementioned issues, we devise an approach for the NLI problem in the medical domain that is equipped with an **adaptive encoding scheme** to integrate context-relevant domain knowledge into the text representation more effectively. In particular, a mixture model is employed to adaptively leverage supplementary external resources, that provide contextual evidence to disambiguate the concept sense for each word, and thus facilitate in learning more semantically refined text embeddings, as well as, account for the missing words. Furthermore, in order to infuse important inferential clues between the premise-hypothesis tokens, context-relevant relational embeddings elicited from knowledge graph is encoded through **multi-source attention mechanism**. We dub the proposed framework as **Mu**lti-**S**ource **K**nowledge **A**daptive Inference **N**etwork (**MUSKAN**).

2 Related Works

2.1 Natural Language Inference

Natural Language Inference lies at the core of many NLP problems (Harabagiu and Hickl, 2006; Rush et al., 2015; Pasunuru and Bansal, 2017). Recently, deep learning has achieved great success in NLI. Current neural models for NLI can be categorized into two main groups of frameworks (1) *sentence encoding models* and (2) *sentence pair interaction models*, as discussed below:

In the sentence encoding framework, the sentence pair is modeled by encoding each sentence separately and the semantic relationship computed based on their similarity. InferSent (Conneau et al., 2017) first encodes the sentences using a recurrent model and then performs element-wise product and absolute difference to capture the relations between the sentences. A stacked BiLSTM is used in the Gated BiLSTM model proposed by (Chen et al., 2017b), which first applies intra-sentence gated attention[4] to bring the sentences to fixed length vectors, and then relation information similar to InferSent is computed.

Whereas in the case of sentence pair interaction framework, word-level interactions are captured using some sort of alignment mechanism (e.g., attention), which are then aggregated to a fixed-length vector to make the final decision. ESIM (Chen et al., 2016) first uses BiLSTM to capture sequential context and then models local inference between word pairs using attention; it then enhances them by computing relation information similar to InferSent/Gated BiLSTM but at the word-level, which is then aggregated to fixed length vectors using a second BiLSTM. In addition, it also incorporates syntactic parsing information with a second similar network. Match-LSTM (Wang and Jiang, 2015) first uses LSTMs to encode the sentences, then computes word-by-word matching using an attention scoring function for each time step, where the last

[2] compared to open domain NLI datasets

[3] we call such words "missing words" in this paper

[4] note that this attention does not capture cross-features between corresponding words in the two sentences, so not grouped into the second group

hidden state is used to represent the sentence representation.

2.2 NLI and External Knowledge

Utilizing external knowledge has shown improvement in performance for some NLI works (Chen et al., 2017a; Wang et al., 2019; Li et al., 2019). Knowledge from WordNet (Miller, 1995) is leveraged in work by (Chen et al., 2017a) to enhance the different components of the NLI model. (Kang et al., 2018) uses the hypernym/hyponym information from three different external linguistic resources, namely WordNet, PPDB (Ganitkevitch et al., 2013) and SICK (Marelli et al., 2014), to generate adversarial examples which are used to augment and train the text entailment system in order to make it robust. (Wang et al., 2019) uses WordNet, ConceptNet and DBPedia (Auer et al., 2007) to incorporate knowledge graphs into text-based NLI models.

In medical NLI, external knowledge is provided as domain knowledge that exists in the form of medical ontology or knowledge base. Work by (Jin et al., 2019) incorporates relational information from UMLS into pre-trained BioELMO[5] and BioBERT (Lee et al., 2020) embeddings. (Romanov and Shivade, 2018) similarly uses domain-specific knowledge from UMLS, however, they modify the pre-trained embeddings using retrofitting. They also experiment with knowledge-directed attention in ESIM and InferSent models.

The main drawback of the aforementioned approaches is that they rely on the context-independent domain knowledge returned by UMLS, which either returns an inaccurate mapping or no mapping and hence could possibly lead to wrong inference predictions. This work addresses these drawbacks with competitive performance on the MedNLI dataset.

3 Approach Overview

We treat the task of Natural Language Inference (NLI) as a supervised classification task and state it as follows: given a premise sentence $\mathbf{p} = (w_1^p, ..., w_m^p)$ with length m, a hypothesis sentence $\mathbf{h} = (w_1^h, ..., w_n^h)$ with length n and the corresponding lexical (i.e., UMLS concept) and relational (i.e., UMLS relation triples) domain knowledge for the sentences represented as $\mathbf{c^p} = (c_1^p, ..., c_m^p)$, $\mathbf{c^h} = (c_1^h, ..., c_n^h)$ and $\mathbf{r^p} = (r_1^p, ..., r_m^p)$, $\mathbf{r^h} = (r_1^h, ..., r_n^h)$ respectively, our goal is to learn a classifier $\mathcal{P}$ (a neural network in our case) which is able to predict the inference relation $y \in Y$ between $\mathbf{p}$ and $\mathbf{h}$ by leveraging the domain knowledge, where $Y = \{entailment, contradiction, neutral\}$. *Entailment* means that when **p** is true, then **h** must be true; *contradiction* means when **p** is true, then **h** must be false; *neutral* means neither entailment nor contradiction. More formally,

$$y* = \arg\max_{y \in Y} \mathcal{P}(y|p, h, c^p, c^h, r^p, r^h) \qquad (1)$$

Here, $c_i^p$and c_j^h are the i-th and j-th concept, and r_i^p and r_j^h are the i-th and j-th relation triple of the premise and hypothesis respectively. Note that a word w_i^p/w_j^h in the premise/hypothesis could also be an abbreviation, which here we collectively call word.

As aforementioned, although incorporating domain knowledge from the UMLS helps to understand medical semantics in the text that go beyond basic linguistic understanding, it could also aggravate the inference process due to the missing words and the inaccurate mappings. To this end, we supplement the UMLS with other external resources in order to soft-align the context-relevant domain information to each word in the text. Figure 2 illustrates a high-level overview of the architecture of our proposed model, **MUSKAN**. It follows the encode-match-classify framework of general text-based NLI models (Chen et al., 2016; Parikh et al., 2016). In the encoder layer, an **adaptive encoding scheme** encodes each word in the premise/hypothesis sentence by integrating *refined* concept embeddings into the text representation, where the refinement is done by leveraging contextual evidence from the supplementary external resources. Then in the matching layer, the adaptive lexical encodings are enhanced with refined relational information codified in knowledge graphs using **multi-source attention**, that facilitates in semantically aligning and aggregating the interactions between the premise-hypothesis words. Finally, the classification layer composes the pair of sentences to a fixed length vector and predicts their relation. More details of each component will be presented in the next sections.

3.1 Adaptive Lexical Encoding

For accurately capturing the relevant lexical semantics of a word in its context, the adaptive encoding

[5] https://github.com/Andy-jqa/bioelmo

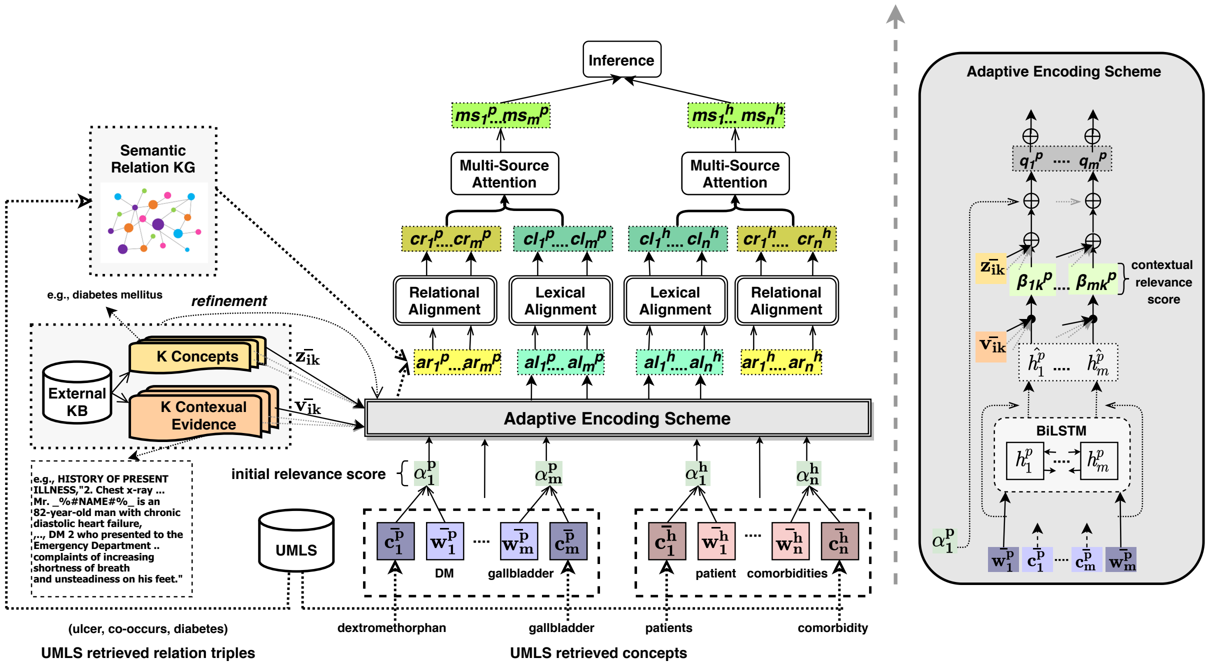

Figure 2: Architecture of our proposed MUSKAN model. The overall framework is shown to the left in a bottom-up fashion, with the sample pair in Figure 1 as the input. The Adaptive Encoding Scheme is illustrated in detail in the figure on the right, taking the premise as an example input.

scheme exploits other external resources alongside the UMLS. A mixture model similar to (Yang and Mitchell, 2019) is employed to refine the initial UMLS concept embedding with a weighted sum of candidate concept vectors, where the weights are adaptively adjusted based on the relevance of the concept's supporting evidence to the word context. The refined concept embeddings are then integrated into the respective text representations to output the final encoded word representation.

To be specific, given the premise $\mathbf{p} = (w_1^p, ..., w_m^p)$ and the corresponding UMLS concepts $\mathbf{c^p} = (c_1^p, ..., c_m^p)$ (the same procedure is utilized for the $\mathbf{h}$ and $\mathbf{c^h}$ hypothesis pair, but for ease of presentation, we only describe the adaptive lexical encoding for the premise; subsequently, we drop the superscript p), each word is first converted to a d-dimensional vector using a pre-trained word embedding method to yield the embedded representations $\bar{\mathbf{p}} = (\bar{w}_1, ..., \bar{w}_m)$ and $\bar{\mathbf{c}} = (\bar{c}_1, ..., \bar{c}_m)$ respectively [6]. We then compute the initial relevance score, α_i, between the i-th word and its concept to get an idea of the degree to which the UMLS-retrieved lexical knowledge is useful in distilling the semantic meaning of the current word. It is computed as,

$$\alpha_i = \bar{w_i}^\top W_a \bar{c_i} \tag{2}$$

where W_a is a trainable weight matrix. However, as the UMLS returns mappings without considering the context in which the word occurs, it is possible that the retrieved concept expresses rather a wrong meaning, which could mislead the inference process. For example, consider the term “DM" in Figure 1; “dextromethorphan" is returned as the matching concept by the UMLS [7], but “diabetes mellitus" is actually the correct concept in that specific context. Evidently, this wrong domain knowledge can avert the model from establishing important inferential clues like the semantic similarity between “diabetes mellitus" in the premise and “diabetes" in the hypothesis, which would otherwise help it to conclude in a conclusive manner that the semantic relationship is entailment. Besides, the UMLS does not offer total coverage of concepts across the whole natural language, which means that for some medical domain-specific jargon, such as the abbreviations “SFA" and “MDR", there exists no corresponding concepts.

To tackle these issues, we resort to external resources that can provide supporting evidence to val-

[6] note that in the case of missing words without any concept/relation mapping, we provide a synthetic placeholder and set its embedding to zero

[7] based on the highest MetaMap Indexing (MMI) score

idate the relevant domain knowledge in the right textual context. Concretely, for the i-th word in $\mathbf{p}$, a set of K candidate concepts, $z_{i1}, ..z_{iK}$, related to it and their corresponding supporting evidences, $v_{i1}, ..v_{iK}$, are first retrieved (discussed in Sections 3.1.1 and 3.1.2). The candidate concepts and their supporting evidences are then embedded as d-dimensional vectors, $\bar{\mathbf{z}}_{\mathbf{i}} = (\bar{z_{i1}}, ..., \bar{z_{iK}})$ and $\bar{\mathbf{v}}_{\mathbf{i}} = (\bar{v_{i1}}, ..., \bar{v_{iK}})$ respectively, using the same pre-trained embedding method as discussed before. LSTM (Hochreiter and Schmidhuber, 1997) is a special variant of Recurrent Neural Networks (Williams and Zipser, 1989) and has shown to capture long-range dependencies and nonlinear dynamics between words. In order to model the contextual information of each word that is indicative of its semantic meaning, we use a BiLSTM that processes the premise $\mathbf{p}$ in both forward and backward directions and produces the hidden states $\hbar = \{\hbar_1, ..., \hbar_m\}$. Subsequently, for the i-th word, its context vector $\hat{\hbar}_i \in \mathbb{R}^{2\mathrm{h}}$ is computed as:

$$\hat{\hbar}_i = \sigma(W_b \hbar_i + W_c \bar{w}_i) \qquad (3)$$

where W_b and W_c are weight matrices to be learned, $\mathtt{h}$ is the number of hidden units and σ indicates the sigmoid function. In order to gauge the suitability of each candidate concept, z_{ik} where $k \in [1, K]$, as a more semantically similar concept to the word compared to the initially retrieved context-independent UMLS concept c_i, we compute the relevance of its embedded supporting evidence to the current word context as:

$$\beta_{ik} = \hat{\hbar}_i^{\top} W_d \bar{v_{ik}} \qquad (4)$$

where W_d is a trainable weight matrix. The initial concept embedding is, henceforth, refined with a mixture model that is formulated as a weighted sum of the candidate concept vectors, where the weights are the relevance scores. The *refined lexical knowledge vector*, $q_i \in \mathbb{R}^d$, is defined as:

$$q_i = \alpha_i \bar{c}_i + \sum_{k=1}^{K} \beta_{ik} \bar{z_{ik}} \qquad (5)$$

Here, $\alpha_i + \sum_{k=1}^{K} \beta_{ik} = 1$ to ensure that the weights are adaptively adjusted according to the concepts' relevance to the word context, and apparently the contribution from the most relevant concept will be properly emphasized with a higher relevance score. In the case of missing words which have no corresponding UMLS concepts and hence make the first term in equation 5 zero, the candidate concept vectors retrieved from external resource will compensate for that through the second term.

Finally, the refined knowledge vector is integrated into its original contextual representation to get the *adaptive lexical embedding*:

$$al_i = \hbar_i + q_i \qquad (6)$$

We consider al_i as the final representation of the i-th word that results in the encoded premise $\mathbf{al^p} = (al_1^p, ..., al_m^p)$ (similarly for the encoded hypothesis $\mathbf{al^h} = (al_1^h, ..., al_n^h)$), which are passed as inputs into the next component.

3.1.1 Candidate Concepts

To select K candidate concepts for each concept, we measure the relevance between the respective contextual evidence (collected as described in Section 3.1.2) by performing dot product between their embeddings. For each abbreviation, its possible expansions in the Abbreviation Sense Inventory dataset are considered as the candidate concepts. While for each word (non-abbreviation), the candidates are selected from the total concept space (~5300 medical concepts). We set K to 5 based on hyperparameter analysis on the validation set.

3.1.2 Contextual Evidence

The contextual evidence for each word/abbreviation in the medical text is collected as snippet of clinical note from two different external resources respectively. For each word (non-abbreviation) in the text, we leverage the clinical notes in the MIMIC-III critical care dataset (Johnson et al., 2016) to extract the relevant snippet in which the word appears. While for abbreviations, we first check against the more specialized Clinical Abbreviation Sense Inventory dataset (Moon et al., 2012). It contains 440 most frequently used abbreviations selected from 352,267 dictated clinical notes. Each abbreviation instance is annotated with its long form, the source sentence where the abbreviation appears, along with other information. The source sentence is fed as the contextual evidence for the abbreviation. If it happens that the abbreviation is not found in the specialized dataset, then we resort to the MIMIC-III critical care dataset.

3.2 Matching with Multi-Source Attention

In order to capture fine-grained word-level information for semantic comparisons that lead to improved local inferential decisions, our proposed model attends over the word pair interactions between the encoded premise $\mathbf{al^p}$ and the encoded hypothesis $\mathbf{al^h}$ at both the lexical and relational levels using **multi-source attention mechanism**. Figure 1 depicts the motivations for introducing this scheme. At the lexical level words are aligned to model their semantic similarity (i.e., in red), while the relational alignment reveals the innate semantic relations existing between medical entities (i.e., in green). This fine-grained alignment simulated by the multi-source attention is important for medical NLI as the semantic relation between the premise-hypothesis depends largely on the relations of aligned semantic units, which in turn require reasoning over a range of domain-specific knowledge phenomena.

The adaptive encodings outputted from the previous component already capture the lexical semantics appropriately, so *lexical alignment* soft-aligns the adaptive representations of the i-th word in the encoded premise $\mathbf{al^p}$ and the j-th word in the encoded hypothesis $\mathbf{al^h}$ into an alignment matrix L $\in \mathbb{R}^{m \times n}$. It is calculated as:

$$l_{ij} = al_i^{p\top} \cdot al_j^h \tag{7}$$

Using these cross-sentence word attention weights, the *lexical context vector*, $\mathbf{cl^p_i}$, of the i-th word in the encoded premise is computed to characterize the most semantically similar parts in the encoded hypothesis and vice versa:

$$\gamma_{ij} = \frac{exp(l_{ij})}{\sum\limits_{k=1}^{n} exp(l_{ik})}, \quad cl_i^p = \sum_{j=1}^{n} \gamma_{ij} al_i^p \tag{8}$$

$$\delta_{ij} = \frac{exp(l_{ij})}{\sum\limits_{k=1}^{m} exp(l_{kj})}, \quad cl_j^h = \sum_{i=1}^{m} \delta_{ij} al_i^h \tag{9}$$

As for *relational alignment*, first the knowledge graph for each word - summarizing its relationships with other concepts in the medical domain - is retrieved from the UMLS (next sub-section). It then converts them to adaptive relational embeddings ar^p/ar^h with a graph representation technique, which are attended over the same way as the lexical alignment (al^p/al^h in equations 7, 8, 9 replaced with ar^p/ar^h), but for modeling the explicit dependency relationship between the word graph representations to produce the *relational context vectors*, $\mathbf{cr^p}$ and $\mathbf{cr^h}$, for the premise and hypothesis respectively.

The interactive features in the lexical context vector and the relational context vector are then merged as the *multi-source context vector*:

$$ms_i^p = W_{m1}([cl_i^p; cr_i^p]) + b_{m1} \tag{10}$$

$$ms_j^h = W_{m2}([cl_j^h; cr_j^h]) + b_{m2} \tag{11}$$

where W_{m1} and W_{m2} are trainable weight matrices and [;] indicates concatenation.

3.2.1 Adaptive Embedding of Relational Knowledge

The relational information between medical concepts can provide invaluable inferential clues to enhance the interactive features between the word pairs in the sentences. In order to create the relational knowledge graph, we resort to the Semantic Network within the UMLS. We first use MetaMap to map the words/phrases of the premise-hypothesis pairs in the MedNLI dataset to their corresponding UMLS concepts. This gives us a total of $\sim$ 5300 unique medical concepts, which form the nodes of the knowledge graph. Two medical concepts form an edge if there exists a relationship between their respective semantic types in the Semantic Network and we get a total of $\sim$ 15,000,000 edges.

We employ graph attention (Veličković et al., 2017; Guan et al., 2019) to represent the knowledge graph as low-dimensional vector(s) for each medical concept(s). In order to propagate the *refined* lexical knowledge into the relational embeddings, we compute a mixture of the graph embeddings between the UMLS retrieved concept and its K candidate concepts, where the same α and β weights from adaptive lexical encoding are used. This way, the graph embedded relational knowledge will align appropriately with the context-aware medical concept. First, for each concept and its candidates, their respective one-hop graph is retrieved from the aforementioned relational knowledge graph. That is, say for the i-th medical concept (similarly for its candidate concepts z_{ik}, where $k \in [1, K]$), its one-hop graph $G(i)$ is represented using its relation triples as $G(i) = \{r_1, ...r_{N_{deg_i}}\}$. Here, the n-th triple indicates semantic relationship of the i-th concept with a neighboring concept and can be written as $(head_n, r_n, tail_n)$, where the i-th concept is the head concept in each. Note that we use the concept's preferred name for each concept, and hence represent

Table 1: Accuracy performance of different models on the development and test sets of MedNLI. We use 768-*d* BioBERT embeddings in all. g/l indicates the percentage gain(+)/loss(-) compared to ESIM w/K.

	GBLM	IS	IS w/K	MLM	ESIM	ESIM w/K	MUSKAN	Ab_1	Ab_2
Dev	73.11	74.02	74.79	74.98	76.37	78.88	**80.09**	76.99	78.13
Dev g/l	-7.31	-6.16	-5.19	-4.94	-3.18	N/A	+1.53	-2.39	-0.95
Test	72.15	73.82	74.14	74.03	75.19	77.26	**79.42**	76.02	77.55
Test g/l	-6.61	-4.45	-4.04	-4.18	-2.68	N/A	+2.79	-1.60	+0.37

all head and tail concepts using the previous pre-trained embedding. Graph attention uses attention mechanism to learn the relative weight between two connected concepts (Wu et al., 2020), that is used to obtain the graph vector, $\hat{g}_i$, as:

$$\hat{g}_i = \sum_{n=1}^{N_{deg_i}} \mu_n[head_n; tail_n] \tag{12}$$

$$\mu_n = \frac{exp(\hat{\mu}_n)}{\sum_{n'=1}^{N_{deg_i}} \hat{\mu}_{n'}} \tag{13}$$

$$\hat{\mu}_n = (W_{r1}rel_n)tanh(W_{r2}head_n + W_{r3}tail_n) \tag{14}$$

where N_{deg_i} is the degree of concept i, and rel_n is a trainable relation vector for relation r_n and is randomly initialized.

The *adaptive relational embedding* is then computed as a mixture model using the graph vectors for the concept and its candidates, as shown below:

$$g_i = \alpha_i\hat{g}_i + \sum_{k=1}^{K} \beta_{ik}\hat{g}_{ik} \tag{15}$$

For notation consistency, we instead use the notations ar_i^p and ar_j^h to denote the adaptive relational embedding of the i-th/j-th concept in the premise and hypothesis respectively.

3.3 Inference

In order to aggregate the inferential semantics at the word level to a sentence representation, we first enrich the context vectors with similarity and closeness information (Chen et al., 2016; Kumar et al., 2016):

$$s_i^p = F([al_i^p; ar_i^p; ms_i^p; al_i^p - ms_i^p; ar_i^p - ms_i^p; al_i^p \odot ms_i^p; ar_i^p \odot ms_i^p]) \tag{16}$$

$$s_j^h = F([al_j^h; ar_j^h; ms_j^h; al_j^h - ms_j^h; ar_j^h - ms_j^h; al_j^h \odot ms_j^h; ar_j^h \odot ms_j^h]) \tag{17}$$

where $F(.)$ is a standard projection layer with ReLU activation function followed by a BiLSTM.

Finally, a pooling layer, comprising max and mean pooling, is used to convert the vectors into a fixed-length vector and then fed into a 2-layer multi-layer perception (MLP) classifier to make the final inference prediction. The entire model is trained end-to-end, through minimizing the cross-entropy loss.

4 Experiments and Results

4.1 Data

We evaluate performance of our model on the only publicly available dataset for this task, namely MedNLI (Romanov and Shivade, 2018). Each instance in this expert-annotated dataset is a premise-hypothesis pair, along with a gold label indicating their inferential relationship. The training, development and test sets consist of 11,232, 1395 and 1422 sentence pairs respectively.

4.2 Baselines

We compare our model against both sentence encoding-based (InferSent (IS) and Gated BiLSTM (GBLM)) and sentence pair interaction-based (ESIM and Match-LSTM (MLM)) baselines. Furthermore, we incorporate domain knowledge in the form of UMLS medical concepts and relation information into the best performing model from each group (i.e., InferSent and ESIM). In the case of InferSent, the knowledge features are fed during encoding into the text representation; for ESIM, we also incorporate it into the attention. We refer to these knowledge-enhanced versions of the baselines with the “w/K" suffix.

4.3 Implementation Details

We use pre-trained 768-*d* BioBERT (Lee et al., 2020) vectors to initialize all word and concept embeddings in the adaptive lexical encoding step, with update during training. The hidden states of both the BiLSTMs during encoding and inference are set to 384. An Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 0.0005 is used to optimize all the trainable weights. The mini-batch size is set to 64.

4.4 Results

Table 1 reports the accuracy of the models on the development and test sets of the MedNLI dataset. MUSKAN outperforms all the baselines by a significant margin with a test accuracy of 79.42%. Specifically, there is a 2.79% performance improvement in comparison to the best performing baseline, ESIM w/K. Although ESIM w/K exploits the semantic knowledge in UMLS, we can assert that refining this knowledge using an adaptive encoding scheme based on contextual evidence is able to alleviate the noise introduced by the domain knowledge, and hence leads to a major boost in performance.

A general observation is that the sentence encoding baselines perform poorly compared to the counterpart sentence interaction ones. The main limitation of the encoding approaches is that they fail to capture the interactions between the premise-hypothesis words, that could otherwise provide important alignment information for inference.

To further ascertain the effectiveness of our model, we evaluate the contributions of key factors in our method by performing an ablation study. The ablated versions of our model are shown on the far right of Table 1 as Ab_1 and Ab_2. Since our proposed method encodes the premise-hypothesis sentences by integrating the *refined* domain knowledge into the text representation, we wonder how the model would perform without this adaptive encoding. So in the encoding step of Ab_1, we concatenate the initial UMLS retrieved concept embedding to the corresponding text representation, which is then passed as input to the subsequent component. We observe that this leads to a drop in performance by 4.28% compared to the whole model. This verifies our intuition that embedding the context-relevant domain knowledge can indeed improve understanding the semantics of the text. In the case of Ab_2, we use just the adaptive lexical encoding to compute the attention matrix, and can see that this declines the performance by 2.35%. This shows that our proposed model works more effectively by capturing the cross-features from both adaptive lexical and adaptive relational representations at the same time using multi-source attention.

4.5 Case Study

There are two different visualizations to demonstrate our model's interpretability. First, the visualization of *lexical alignment* shows how adaptive lexical encoding helps to align the semantically similar words in the premise-hypothesis sentence pair. Next, the *multi-source attention* visualization enhances the lexical alignment by highlighting the salient words that well represent the semantic relation between them.

The sub-figures in Figure 3 depict the attention heatmaps yielded by the best performing baseline, ESIM w/K, and our proposed MUSKAN. The darker shade indicates higher importance in classification. The alignment of the words “feeling", “fatigued", “light", and “headed" in p to “weakness" in h is critical in deciding if the former entails the latter. From the highlighted words in the middle sub-figure in Figure 3 for lexical alignment matrix of our proposed model, it can be seen that integrating context-aware medical concepts into the text representation is in fact able to capture this semantics. The abbreviation “USOH" stands for “usual state of health" and expresses a transition from normal to a deterioration of patient's health in this context. In the right sub-figure for multi-source attention matrix, the higher attention put on the words “onset", “prior", “to", “admission", “started" and PCP", and their alignment with “new" are able to capture this nuance. We hypothesize that this is facilitated by the semantic relation information between the medical concepts incorporated through the multi-source attention. On the other hand, from the left sub-figure, it can be seen that ESIM w/K fails to model these context-aware lexical and relational associations due to missing words (e.g., USOH) and inaccurate mappings (e.g., PCP), which result in a wrong prediction.

4.6 Error Analysis

We perform error analysis on the result of MUSKAN which divulges open challenges and directions towards pending future research in medical NLI. Typical errors made by our approach include:

Numeric values: For some premises, the text can describe clinical measurements as numeric val-

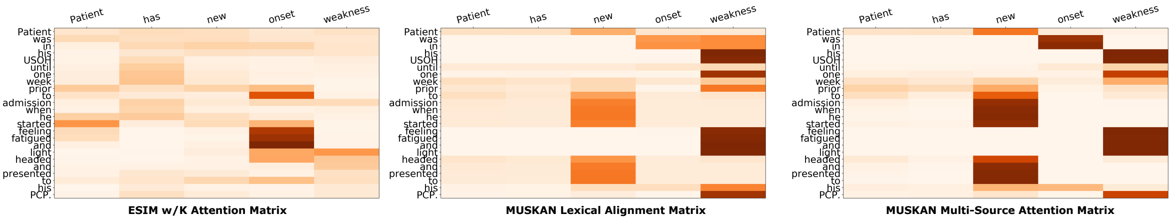

Figure 3: Visualizations of the attention heatmaps for the following instance from the test set of the MedNLI dataset: {p: Patient was in his USOH until one week prior to admission when he started feeling fatigued and light headed and presented to his PCP. h: Patient has new onset weakness. y: Entailment}.

ues, which make it difficult for the model to semantically relate these to the condition conveyed in the hypothesis. For example, looking at the premise in Figure 4a, we can see that the different vital signs, represented with the abbreviation “VS", are expressed in terms of numeric values (e.g., T 98.9, HR 73, BP 121/90). However, for the model to infer that these values indicate that the patient is “hemodynamically stable" is challenging. Hypothesizing, we attribute this fail to the fact that medical notes lack in covering such knowledge and, perhaps, leveraging other external resources such as the Wikipedia or the laboratory test results available in electronic health records (EHR) might help to mitigate this drawback.

Ambiguity: Some instances in the dataset contain words/phrases used in everyday conversation, which could appear as vague terms with respect to medical perspective and result in misclassification. As an example,“handfuls" in the premise in Figure 4b is actually referring to “more medications than directed" in the hypothesis and is an “entailment". However, the ambiguity here lies in that “overdose of Dilaudid" (which has label “neutral") possibly expresses similar concept, and hence leads to a false positive.

Premise: *In the ED, initial VS revealed T 98.9, HR 73, BP 121/90, RR 15, O2 sat 98% on RA.*

Hypothesis: *The patient is hemodynamically stable*

Label: *Entailment*

(a)

Premise: *Today she got into an argument with her husband and felt that she \"wanted to sleep\" and therefore took \"handfuls\" of dilaudid.*

Hypothesis: *She took more medication than directed*

Label: *Entailment*

(b)

Figure 4: Samples from MedNLI dataset to demonstrate error analysis for (a) Numeric values and (b) Ambiguity

5 Conclusion and Future Work

This work discloses the effectiveness of context-aware domain knowledge in medical NLI and proposes a systematic approach to infuse such knowledge using an adaptive encoding scheme. By employing a multi-source attention mechanism that is able to model both the lexical and relational semantics, it is able to mitigate the noise introduced by the abbreviation-like jargon prevalent in medical text. Through both qualitative and quantitative analysis, our proposed framework advances the limited work so far done on medical NLI.

There are several possible directions that could be explored as future work. Firstly, it would be interesting to investigate if enriching the refined domain knowledge with explicit syntactic information (e.g., parse tree) of the premise-hypothesis is helpful. Secondly, we could extract knowledge from other relevant medical knowledge bases and incorporate deeper subgraph information (e.g., two hops). Furthermore, we could test the utility of the proposed framework on downstream NLP applications that similarly suffer from small data size.

6 Acknowledgments

We would like to thank the anonymous reviewers for their careful reading of our manuscript and providing helpful comments and suggestions. This work is supported in part by NSF under grants III-1763325, III-1909323, and SaTC-1930941.

References

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Chen, Q., Zhu, X., Ling, Z., Wei, S., Jiang, H., and Inkpen, D. (2016). Enhanced lstm for natural language inference. *arXiv preprint arXiv:1609.06038*.

Chen, Q., Zhu, X., Ling, Z.-H., Inkpen, D., and Wei, S. (2017a). Neural natural language inference models enhanced with external knowledge. *arXiv preprint arXiv:1711.04289*.

Chen, Q., Zhu, X., Ling, Z.-H., Wei, S., Jiang, H., and Inkpen, D. (2017b). Recurrent neural network-based sentence encoder with gated attention for natural language inference. *arXiv preprint arXiv:1708.01353*.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2013). Ppdb: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764.

Guan, J., Wang, Y., and Huang, M. (2019). Story ending generation with incremental encoding and commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6473–6480.

Harabagiu, S. and Hickl, A. (2006). Methods for using textual entailment in open-domain question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 905–912. Association for Computational Linguistics.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jin, Q., Dhingra, B., Cohen, W. W., and Lu, X. (2019). Probing biomedical embeddings from language models. *arXiv preprint arXiv:1904.02181*.

Johnson, A. E., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

Kang, D., Khot, T., Sabharwal, A., and Hovy, E. (2018). Adventure: Adversarial training for textual entailment with knowledge-guided examples. *arXiv preprint arXiv:1805.04680*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Zhong, V., Paulus, R., and Socher, R. (2016). Ask me anything: Dynamic memory networks for natural language processing. In *International conference on machine learning*, pages 1378–1387.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Li, T., Zhu, X., Liu, Q., Chen, Q., Chen, Z., and Wei, S. (2019). Several experiments on investigating pretraining and knowledge-enhanced models for natural language inference. *arXiv preprint arXiv:1904.12104*.

Lloret, E., Ferrández, O., Munoz, R., and Palomar, M. (2008). A text summarization approach under the influence of textual entailment. In *NLPCS*, pages 22–31.

Long, T., Bengio, E., Lowe, R., Cheung, J. C. K., and Precup, D. (2017). World knowledge for reading comprehension: Rare entity prediction with hierarchical lstms using external descriptions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 825–834.

MacCartney, B. and Manning, C. D. (2009). *Natural language inference*. Citeseer.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., Zamparelli, R., et al. (2014). A sick cure for the evaluation of compositional distributional semantic models. In *LREC*, pages 216–223.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Moon, S., Pakhomov, S., and Melton, G. (2012). Clinical abbreviation sense inventory.

Parikh, A. P., Täckström, O., Das, D., and Uszkoreit, J. (2016). A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*.

Pasunuru, R. and Bansal, M. (2017). Multi-task video captioning with video and entailment generation. *arXiv preprint arXiv:1704.07489*.

Romanov, A. and Shivade, C. (2018). Lessons from natural language inference in the clinical domain. *arXiv preprint arXiv:1808.06752*.

Rush, A. M., Chopra, S., and Weston, J. (2015). A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Wang, S. and Jiang, J. (2015). Learning natural language inference with lstm. *arXiv preprint arXiv:1512.08849*.

Wang, X., Kapanipathi, P., Musa, R., Yu, M., Talamadupula, K., Abdelaziz, I., Chang, M., Fokoue, A., Makni, B., Mattei, N., et al. (2019). Improving natural language inference using external knowledge in the science questions domain. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7208–7215.

Weissenborn, D., Kočiskỳ, T., and Dyer, C. (2017). Dynamic integration of background knowledge in neural nlu systems. *arXiv preprint arXiv:1706.02596*.

Williams, A., Nangia, N., and Bowman, S. R. (2017). A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Williams, R. J. and Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Philip, S. Y. (2020). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*.

Yang, B. and Mitchell, T. (2019). Leveraging knowledge bases in lstms for improving machine reading. *arXiv preprint arXiv:1902.09091*.