Lessons from Natural Language Inference in the Clinical Domain

Alexey Romanov
Department of Computer Science
University of Massachusetts Lowell*
Lowell, MA 01854
aromanov@cs.uml.edu

Chaitanya Shivade
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120
cshivade@us.ibm.com

Abstract

State of the art models using deep neural networks have become very good in learning an accurate mapping from inputs to outputs. However, they still lack generalization capabilities in conditions that differ from the ones encountered during training. This is even more challenging in specialized, and knowledge intensive domains, where training data is limited. To address this gap, we introduce MedNLI[1] – a dataset annotated by doctors, performing a natural language inference task (NLI), grounded in the medical history of patients. We present strategies to: 1) leverage transfer learning using datasets from the open domain, (e.g. SNLI) and 2) incorporate domain knowledge from external data and lexical sources (e.g. medical terminologies). Our results demonstrate performance gains using both strategies.

1 Introduction

Natural language inference (NLI) is the task of determining whether a given *hypothesis* can be inferred from a given *premise*. This task, formerly known as recognizing textual entailment (RTE) (Dagan et al., 2006) has long been a popular task among researchers. Moreover, contribution of datasets from past shared tasks (Dagan et al., 2009), and recent research (Bowman et al., 2015; Williams et al., 2018) have pushed the boundaries for this seemingly simple, but challenging problem.

The Stanford Natural Language Inference (SNLI) dataset (Bowman et al., 2015) is a large, high quality dataset and serves as a benchmark to evaluate NLI systems. However, it is restricted to a single text genre (Flickr image captions) and mostly consists of short and simple sentences. The MultiNLI corpus (Williams et al., 2018) which introduced NLI corpora from multiple genres (e.g. fiction, travel) was a welcome step towards addressing these limitations. MultiNLI offers diversity in linguistic phenomena, which makes it more challenging.

Following these efforts, we explore the problem of NLI in the clinical domain. Language inference in specialized domains such as medicine is extremely complex and remains unexplored by the machine learning community. Moreover, since this domain has a distinct sublanguage (Friedman et al., 2002), clinical text also presents unique challenges (abbreviations, inconsistent punctuation, misspellings, etc.) that differentiate it from open-domain data (Meystre et al., 2008).

In this paper, we address these gaps and make the following contributions:

- Introduce MedNLI - a new, publicly available, expert annotated dataset for NLI in the clinical domain.
- A systematic comparison of several state-of-the-art open domain models on MedNLI.
- A study of transfer learning techniques from the open domain to the clinical domain.
- Techniques for incorporating domain-specific knowledge from knowledge bases (KB) and domain specific data into neural networks.

2 The MedNLI dataset

Let us recall the procedure followed for creating the SNLI dataset: annotators were presented with captions for a Flickr photo (the *premise*) without the photos themselves. They were asked to write three sentences (*hypotheses*): 1) A clearly true description of the photo, 2) A clearly false description, and 3) A description that might be true or false. This procedure produces three training pairs

* Work done during an internship at IBM Research
[1] https://jgc128.github.io/mednli/

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1586–1596
Brussels, Belgium, October 31 - November 4, 2018.

#	Premise	Hypothesis	Label
1	ALT , AST , and lactate were elevated as noted above	patient has abnormal lfts	entailment
2	Chest x-ray showed mild congestive heart failure	The patient complains of cough	neutral
3	During hospitalization , patient became progressively more dyspnic requiring BiPAP and then a NRB	The patient is on room air	contradiction
4	She was not able to speak , but appeared to comprehend well	Patient had aphasia	entailment
5	T1DM : x 7yrs , h/o DKA x 6 attributed to poor medication compliance , last A1c [** 3-23 **] : 13.3 % 2	The patient maintains strict glucose control	contradiction
6	Had an ultimately negative esophagogastroduodenoscopy and colonoscopy	Patient has no pain	neutral
7	Aorta is mildly tortuous and calcified .	the aorta is normal	contradiction

Table 1: Examples from the development set of MedNLI

of sentences for each premise with three different labels: entailment, contradiction, and neutral, respectively. In order to produce a comparable dataset, we used the same approach adjusted for the clinical domain.

2.1 Premise sampling and hypothesis generation

As the source of *premise* sentences, we used the MIMIC-III v1.3 (Johnson et al., 2016) database. With de-identified records of 38,597 patients, it is the largest repository of publicly available clinical data. Along with medications, lab values, vital signs, etc. MIMIC-III contains 2,078,705 clinical notes written by healthcare professionals in English. The *hypothesis* sentences were generated by clinicians.

Clinical notes are typically organized into sections such as `Chief Complaint`, `Past Medical History`, `Physical Exam`, `Impression`, etc. These sections can be easily identified since the formatting for associated section headers often resembles capital letters, followed by a colon. The clinicians in our team suggested `Past Medical History` to be the most informative section of a clinical note, from which critical inferences can be drawn about the patient.

Therefore, we segmented these notes into sections using a simple rule based program capturing the formatting of these section headers. We extracted the `Past Medical History` section and used a sentence splitter trained on biomedical articles (Lingpipe, 2008) to get a pool of candidate *premises*. We then randomly sampled a subset from these candidates and presented them to the clinicians for annotation. Figure 1 shows the exact prompt shown to the clinicians for the annotation task. SNLI annotations are grounded since they are associated with captions of the same image. We seek to achieve the same goal by grounding the annotations against the medical history of the same patient.

> You will be shown a sentence from the `Past Medical History` section of a de-identified clinical note. Using only this sentence, your knowledge about the field of medicine, and common sense:
>
> - Write one alternate sentence that is **definitely** a **true** description of the patient. Example, for the sentence "Patient has type II diabetes" you could write "Patient suffers from a chronic condition"
> - Write one alternate sentence that **might be** a **true** description of the patient. Example, for the sentence "Patient has type II diabetes" you could write "Patient has hypertension"
> - Write one sentence that is **definitely** a **false** description of the patient. Example, for the sentence "Patient has type II diabetes" you could write "The patient's insulin levels are normal without any medications."

Figure 1: Annotation prompt shown to clinicians

As discussed earlier, examples shown in Table 1 depict unique challenges that involve reasoning over domain-specific knowledge. For instance, the first three examples require the knowledge about clinical terminology. The fourth example requires awareness of medications and the last example elicits knowledge about radiology images. We make the MedNLI dataset available[2] through the

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

MIMIC-III derived data repository. Thus, any individual certified to access MIMIC-III can also access MedNLI.

2.2 Annotation collection

Conclusions in the clinical domain are known to be context dependent and a source of multiple uncertainties (Han et al., 2011). We had to ensure that such subjective interpretations do not result in annotation conflicts affecting the quality of the dataset. To ensure agreement, we worked with clinicians and generated annotation guidelines for a pilot study. Two board certified radiologists worked on the annotation task, and were presented with the 100 unique premises each.

Some premises, often marred by de-identification artifacts, did not contain any information from which useful inferences could be drawn, e.g. `This was at the end of [**Month (only) 1702**] of this year.` Such sentences were deemed as invalid for the task and discarded based on clinician judgment. The MIMIC-III dataset contains many de-identification artifacts associated with dates and names (persons and places) which also makes MedNLI more challenging.

After discarding 16 premises, the result of hypothesis generation was a set of 552 pairs. To calculate agreement, we presented pairs generated by one clinician, and sought annotations from the other clinician, determining if the inference was "Definitely true", "Maybe true", or "Definitely false" (Bowman et al., 2015). Comparison of these annotations resulted in a Cohen's kappa of $\kappa = 0.78$. While this is substantial if not perfect agreement by itself (McHugh, 2012), it is particularly good given the challenging nature of NLI and the complexity of the domain.[3]

On reviewing the annotations, we found that labeling differences between "Definitely true" and "Maybe true" were the major source of disagreement. This was primarily because one clinician would think of a scenario that is generally true, while the other would think of assumptions (e.g. patient might be lying, or patient might be pregnant) when it would not.

A discussion with clinicians concluded that the annotation guideline was clear and any person with a formal background of medicine should be able to complete the task successfully. To generate the final dataset, we recruited two additional clinicians, both board certified medical students pursuing their residency programs. Unlike SNLI, we did not collect multiple annotations per sentence pair because of the time and funding constraints.

2.3 Dataset statistics

Together, the four clinicians worked on a total of 4,683 premises over a period of six weeks. The resulting dataset consists of 14,049 unique sentence pairs. Following Bowman et al. (2015), we split the dataset into training, development, and testing subsets and ensured that no premise was overlapping between the three subsets. Table 2 presents key statistics of MedNLI.

Dataset size	
Training pairs	11232
Development pairs	1395
Test pairs	1422
Average sentence length in tokens	
Premise	20.0
Hypothesis	5.8
Maximum sentence length in tokens	
Premise	202
Hypothesis	20

Table 2: Key statistics of the dataset

3 Models

To establish a baseline performance on MedNLI, we experimented with a feature-based system. To further explore the performance of modern neural networks-based systems, we experimented several models of various degrees of complexity: Bag of Words (BOW), InferSent (Conneau et al., 2017) and ESIM (Chen et al., 2017). Note that our goal here is not to outperform existing models, but to explore the relative gain of the proposed methods, and compare them to a baseline. We used the same set of hyperparameters in all models to ensure that any difference in performance is exclusively due to the algorithms.

Feature-based system We used a gradient boosting classifier incorporating a variety of hand crafted features. Apart from standard NLP features, we also infused clinical knowledge from the Unified Medical Language System (UMLS) (Bodenreider, 2004). Each terminology in the UMLS can be viewed as a graph where nodes represent medical concepts, and edges represent relations

[3] Rajpurkar et al. (2017) report F1 < 0.45 for four radiologists when compared among themselves

between them. These are canonical relationships found in ontologies such as *IS A* and *SYNONYMY*. For instance, *diabetes IS A disorder of the endocrine system*. The domain specific features we added to the model represent similarity between UMLS concepts from the premise and the hypothesis, based how close they appear in the UMLS graph (Pedersen et al., 2007). Following (Shivade et al., 2015; Pedersen et al., 2007) we used the SNOMED-CT terminology in our experiments.

The groups below summarize the feature sets used in our model (35 features in total):

1. BLEU score
2. Number of tokens (e.g. min, max, difference)
3. Negations (e.g. keywords such as *no, do not*)
4. TF-IDF similarity (e.g. cosine, euclidean)
5. Edit distances (e.g. Levenshtein)
6. Embedding similarity (e.g. cosine, euclidean)
7. UMLS similarity features (e.g. shortest path distance between UMLS concepts)

Bag of words We use a bag-of-words (BOW) model as a simple baseline for the NLI task: the *Sum of words* model by Bowman et al. (2015) with a small modification. While Bowman et al. (2015) use $tanh$ as the activation function in the model, we use ReLU, since it trained faster and achieved better results (Glorot et al., 2011). In order to represent an input sentence as a single vector, this architecture simply sums up the vectors of individual tokens. The premise and hypothesis vectors are then concatenated and passed through a multi-layer neural network. Recent work shows that even this straightforward approach encodes a non-trivial amount of information about the sentence (Adi et al., 2017).

InferSent InferSent (Conneau et al., 2017) is a model for sentence representation that demonstrated close to state-of-the-art performance across a number of tasks in NLP (including NLI) and computer vision. The main differences from the BOW model are as follows:

- A bidirectional LSTM encoder of input sentences and a max-pooling operation over timesteps are used to get a vector for the premise (p) and for the hypothesis (h);
- A more complex scheme of interaction between the vectors p and h to get a single vector z that contains all the information needed to produce a decision about the relationship between the input sentences: $z = [p, h, |p - h|, p * h]$.

ESIM The ESIM model, developed by Chen et al. (2017), is shown in Figure 2. It is a fairly complex model that makes use of two bidirectional LSTM networks. The basic idea of ESIM is as follows:

- The first LSTM produces a sequence of hidden states.
- Pairwise attention matrix e is computed between all tokens in the premise and the hypothesis to produce new sequences of "attended" hidden states, which are then fed into the second LSTM.
- Max and average pooling are performed over the output of the LSTMs.
- The output of the pooling operations is combined in a way similar to the InferSent model.

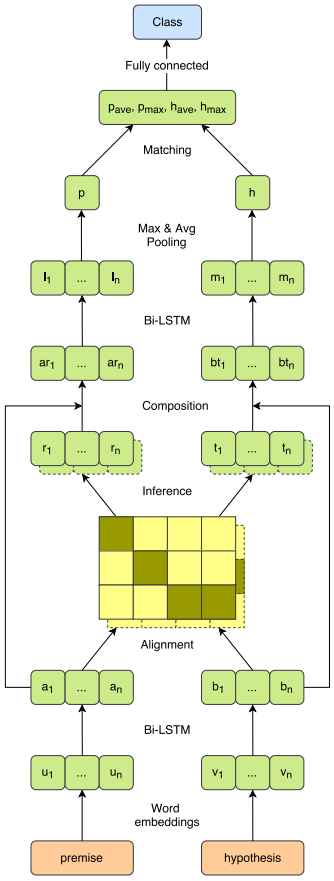

Figure 2: ESIM model. Dashed blocks illustrate the knowledge-directed attention matrix and the corresponding vectors (see Section 4.2.2 for details).

The three aforementioned models exemplify the architectures that are, perhaps, the most widely used for NLI task, spanning from simple bag-of-words approaches to complicated models with Bi-LSTM and inter-sentence attention. We additionally experimented with a plain Bi-LSTM model as well as GRU (Cho et al., 2014), but since their performance was not remarkable (in the same range as BOW) we do not report it here.

4 Transfer learning

Given the existence of larger general-domain NLI datasets such as SNLI and MultiNLI, it stands to reason to try to leverage them to improve the performance in the clinical domain. Transfer learning has been shown to improve performance on variety of tasks such as: machine translation on low-resource languages (Zoph et al., 2016) and also some tasks from the bio-medical domain in particular (Sahu and Anand, 2017; Lee et al., 2018). To see if a corresponding boost would be possible for the NLI task, we investigated three common transfer learning techniques on the MedNLI dataset using SNLI and five different genres from MultiNLI.

Direct transfer is the simplest method of transfer learning. After training a model on a large *source domain* dataset, the model is directly tested on the *target domain* dataset. If the source and the target domains are similar to some extent, one can achieve a reasonable accuracy by simply applying a model pre-trained on the source domain to the target domain. In our case the source domain is general domain in SNLI and the various genres in MultiNLI, and the target domain is clinical.

Sequential transfer is the most widely used technique. After pre-training the model on a large source domain, the model is further fine-tuned using the smaller training data of the target domain. The assumption is that while the model would learn domain-specific features, it would also learn some domain-independent features that will be useful for the target domain. Furthermore, the fine-tuning process would affect the learned features from the source domain and make them more suitable for the target domain.

Multi-target transfer is a more complex method involving separation of the model into three components (or layers):

- *The shared component* is trained on both the source and target domains;
- *The source domain component* is trained only during the pre-training phase and does not participate in the prediction of the target domain;
- *The target domain component* is trained during the fine-tuning stage and it produces the predictions together with the shared component.

The motivation for multi-target transfer is that performance should be improved by splitting deeper layers of the model into domain-specific parts and having a shared block early in the network, where it presumably learns domain-independent features.

4.1 Word embeddings

Another way to improve the accuracy on the target domain is to use domain-specific word embeddings instead of, or, in addition to, open-domain ones. For example, Stanovsky et al. (2017) achieved state of the art results on the task of recognizing Adverse Drug Reaction using graph-based embeddings trained on the "Drugs" and "Diseases" categories from DBpedia (Lehmann et al., 2015), as well as embeddings trained on web-pages categorized as "medical domain".

We experimented with the following publicly available general-domain word embeddings:

- **GloVe$_{[CC]}$**: GloVe embeddings (Pennington et al., 2014), trained on Common Crawl[4].
- **fastText$_{[Wiki]}$**: fastText embeddings (Bojanowski et al., 2017), trained on Wikipedia.
- **fastText$_{[CC]}$**: fastText embeddings, trained on Common Crawl.

Furthermore, we trained fastText embeddings on the following domain-specific corpora:

- **fastText$_{[BioASQ]}$**:A collection of PubMed abstracts from the BioASQ challenge data (Tsatsaronis et al., 2015). This data includes abstracts from 12,834,585 scientific articles from the biomedical domain.
- **fastText$_{[MIMIC-III]}$**: Clinical notes for patients from the MIMIC-III database (Johnson et al., 2016): 2,078,705 notes with 320 tokens in each on average.

Finally, we experimented with initializing word embeddings with pre-trained vectors from general

[4] http://commoncrawl.org/

domain and further training on a domain-specific corpus:

- GloVe$_{[CC]}$ → fastText$_{[BioASQ]}$: GloVe embeddings for initialization, and the BioASQ data for fine-tuning.
- GloVe$_{[CC]}$ → fastText$_{[BioASQ]}$ → fastText$_{[MIMIC-III]}$: GloVe embeddings for initialization, and two consequent fine-tuning using the BioASQ and MIMIC-III data.
- fastText$_{[Wiki]}$ → fastText$_{[MIMIC-III]}$: fastText Wikipedia embeddings for initialization, and the MIMIC-III data for fine-tuning.

Experiments using other approaches to word embeddings, such as `word2vec` (Mikolov et al., 2013) and CoVe (McCann et al., 2017) did not show any gains. All the above trained embeddings are available for download.

4.2 Knowledge integration

Since understanding medical texts requires domain-specific knowledge, we experimented with different ways of incorporating such knowledge into the systems. First, we can modify the input to the system so it carries a portion of clinical information. Second, we can modify the model itself, integrating domain knowledge directly into it.

The UMLS is the largest, publicly available, and regularly updated database of medical terminologies, concepts, and relationships between them. It can be viewed as a graph where clinical concepts are nodes, connected by edges representing relations, such as synonymy, parent-child, etc. Following past work, we restricted to the SNOMED-CT terminology in UMLS and experimented with two techniques for incorporating knowledge: retrofitting and attention.

4.2.1 Retrofitting

Retrofitting (Faruqui et al., 2015) modifies pre-trained word embeddings based on an ontology. The basic idea is to try to bring the representations of the concepts that are connected in the ontology closer to one another in vector space. The authors showed that retrofitting using WordNet (Fellbaum, 1998) synsets improves accuracy on several word-level tasks, as well as sentiment analysis.

4.2.2 Knowledge-directed attention

Attention proved to be a useful technique for many NLP tasks, starting from machine translation (Bahdanau et al., 2015) to parsing (Vinyals et al., 2015) and NLI itself (Parikh et al., 2016; Rocktäschel et al., 2016). In most models (including the ESIM model that we use in our experiments) attention is learned in an end-to-end fashion. However, if we have knowledge about relationships between concepts, we could leverage it to explicitly tell the model to attend to specific concepts during the processing of the input sentence.

For example, there is an edge in SNOMED-CT from the concept *Lung consolidation* to *Pneumonia*. Using this information, during the processing of a sentence pair

- **Premise** The patient has *pneumonia*.
- **Hypothesis** The patient has a *lung* disease.

the model could attend to the token *lung* while processing *pneumonia*.

We propose to integrate this knowledge in a way similar to how attention is used in the ESIM model. Specifically, we calculate the attention matrix $e \in \mathbb{R}^{n \times m}$ between all pairs of tokens a_i and b_j in the inputs sentences, where n is the length of the hypothesis and m is the length of the premise. The value in each cell reflects the length of the shortest path l_{ij} between the corresponding concepts of the premise and the hypothesis in SNOMED-CT (the shorter is the path, the higher is the value).

This process could be informally described as follows: each token $\tilde{a_i}$ of the premise is a weighted sum of relevant tokens b_j of the hypothesis, according to the medical ontology, and vice versa. This enables the medical domain knowledge to be integrated directly into the system.

We used the original tokens a_i as well as the attended $\tilde{a_i}$ inside the model for both InferSent and ESIM. For InferSent, we simply concatenate them across the time dimension:

$$\hat{a} = [a_1, a_2, \ldots, a_n, \tilde{a_1}, \tilde{a_2}, \ldots, \tilde{a_n}]$$

where n is the length of the inputs sequence. For the ESIM model, we concatenate a_i and $\tilde{a_i}$ before passing them to the composition layer (see Figure 2 and Section 3.3 in the original paper (Chen et al., 2017)). This enables the model to learn the relative importance of both the token and the knowledge directed attention.

Set	Features	BOW	InferSent	ESIM
Dev	51.9	71.9	**76.0**	74.4
Test	51.9	70.2	**73.5**	73.1

Table 3: Baseline accuracy on the development and the test set of MedNLI for different models.

5 Results and discussion

We implemented all models using PyTorch[5] and trained them with the Adam optimizer (Kingma and Ba, 2015) until the validation loss showed no improvement for 5 epochs. The epoch with the lowest loss on the validation set was selected for testing. We used the GloVe word embeddings (Pennington et al., 2014) in all experiments, except for subsection 5.3. In all experiments we report the average result of 6 different runs, with the same hyperparameters and different random seeds. Medical concepts in SNOMED-CT were identified in the premise and hypothesis sentences using Metamap (Aronson and Lang, 2010). The code for all experiments is publicly available.[6]

5.1 Baselines

Table 3 shows the baseline results: the performance of a model when trained and tested on the MedNLI dataset. The feature-based system performed the worst. As for neural networks-based systems, the BOW model showed the lowest performance on the both development and test sets. The InferSent model, in contrast, achieved the highest accuracy, despite ESIM outperforming it on SNLI. This could be attributed to the fact that ESIM has twice as many parameters as InferSent, and so InferSent overfits less to the smaller MedNLI dataset.

5.2 Transfer learning

As expected, Table 4 shows that direct transfer is worse than the baseline but is still better than a random baseline of 33.3%. Sequential and multi-target transfer learning, in contrast, yields a considerable gain for all the models. The maximum gain is 2.4%, 0.9%, and 0.3% for the BOW, InferSent, and ESIM models correspondingly.

Second, note that the biggest SNLI domain gave the most boost in only two out of six cases, implying that the size of the domain should not be the most important factor in choosing the source domain for transfer learning. The best accuracy for all the models was obtained with the "slate" domain from MultiNLI corpus with sequential transfer (note, however, that the accuracy of ESIM is actually lower than the baseline accuracy). This is consistent with observations of Williams et al. (2018). Finally, although some domains are better for particular transfer learning methods with particular models, there is no single combination that works for all cases.

5.3 Word embeddings

Table 5 shows that simply using of the embeddings trained on the MIMIC-III notes significantly increases the accuracy for all the models. Furthermore, the InferSent models achieves a 3.1% boost with the fastText Wikipedia embeddings, fine-tuned on the MIMIC-III data. Note that the results $\text{fastText}_{[\text{Wiki}]}$ are worse than the baseline $\text{GloVe}_{[\text{CC}]}$ for all models, which could be due to the source corpus size. However, the results on BioASQ are worse than on MIMIC-III, despite the significantly larger corpus of the BioASQ embeddings. Overall, our experiments show the benefit of domain-specific rather than general-domain word embeddings.

5.4 Knowledge integration

5.4.1 Retrofitting

Table 6 shows that retrofitting only hurts the performance. This is in contrast with the results of the original study, where retrofitting was beneficial not only for word-level tasks but also for tasks such as sentiment analysis (Faruqui et al., 2015). We hypothesize that although WordNet and UMLS are structurally similar, significant differences in the content (Burgun and Bodenreider, 2001) might be the reason for these results. Retrofitting should be more useful when it is used on a WordNet-like database where the main relation is synonymy, and tested on tasks such as word similarity tests or sentiment analysis. The UMLS semantic network is more complex and contains relations that may not be suitable for retrofitting.

Moreover, retrofitting works only on directly related concepts in a knowledge graph (although it might affect, to some extent, indirectly related concepts by transitivity). However, as Figure 3 shows, UMLS contains few training pairs that have such concepts (namely, pairs with a path of

[5] https://pytorch.org/
[6] https://jgc128.github.io/mednli/

Source domain	Direct transfer			Sequential transfer			Multi-target transfer		
	BOW	InferSent	ESIM	BOW	InferSent	ESIM	BOW	InferSent	ESIM
snli	-21.8	**-24.2**	-22.8	1.8	-1.8	-2.5	**2.4**	-2.5	-0.7
fiction	**-21.6**	-25.6	**-21.4**	1.3	0.4	-0.5	1.4	0.1	**0.3**
government	-23.8	-27.2	-26.2	1.0	0.8	-0.7	1.3	0.2	0.2
slate	-23.2	-25.7	-21.6	**1.9**	**0.9**	**-0.2**	1.1	**0.6**	-0.1
telephone	-25.7	-27.3	-25.6	1.7	-0.2	-1.1	1.2	0.4	-0.1
travel	-25.4	-29.1	-23.5	1.6	0.0	-0.7	0.2	-0.3	0.1

Table 4: Absolute gain in accuracy with respect to the baseline (see Table 3) on the MedNLI test set for different transfer learning modes. Bold indicates the best source domain for each model and transfer.

Embeddings	BOW	InferSent	ESIM
$\text{fastText}_{[\text{Wiki}]}$	-3.5	-3.5	-4.4
$\text{fastText}_{[\text{CC}]}$ (600B)	-0.6	1.3	-0.3
$\text{fastText}_{[\text{BioASQ}]}$ (2.3B)	0.5	0.6	0.2
$\text{fastText}_{[\text{MIMIC-III}]}$ (0.8B)	**1.1**	2.3	1.2
$\text{GloVe}_{[\text{CC}]} \rightarrow \text{fastText}_{[\text{BioASQ}]}$	0.2	0.7	1.4
$\text{GloVe}_{[\text{CC}]} \rightarrow \text{fastText}_{[\text{BioASQ}]} \rightarrow \text{fastText}_{[\text{MIMIC-III}]}$	0.9	2.7	**1.8**
$\text{fastText}_{[\text{Wiki}]} \rightarrow \text{fastText}_{[\text{MIMIC-III}]}$	0.1	**3.1**	1.7

Table 5: Absolute gain in accuracy with respect to the baseline ($\text{GloVe}_{[\text{CC}]}$) for different word embeddings (the number in parentheses reflects the number of tokens in the corresponding training corpora).

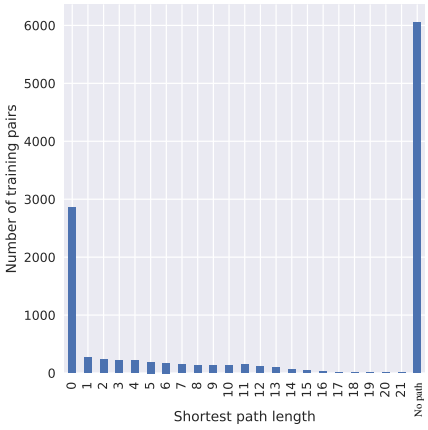

Figure 3: Lengths of the shortest paths between concepts in the premise and the hypothesis. 0 indicates that they contain the same concept.

length 1). In contrast, the lengths of the shortest path in SNLI using WordNet fall close to 1. This suggests that the medical inferences represented in MedNLI requires more complex reasoning, typically involving multiple steps.

As a sanity check, we applied retrofitting to the GloVe embeddings and tested the InferSent model on the "fiction" domain from the MultiNLI corpus. We used the code and lexicons provided by Faruqui et al. (2015) and confirmed that retrofitting hurts the performance in that case as well.

BOW	InferSent	ESIM
-1.7	-2.0	-2.7

Table 6: Absolute gain in accuracy using retrofitting for MedNLI.

5.4.2 Knowledge-directed attention

To evaluate the potential of knowledge-directed attention, let us consider its effect on a baseline embedding ($\text{GloVe}_{[\text{CC}]}$) and a fastText embedding trained on MIMIC-III ($\text{fastText}_{[\text{MIMIC-III}]}$) that showed good performance in section 5.3.

Knowledge-directed attention showed positive effect with the InferSent model on $\text{GloVe}_{[\text{CC}]}$ (0.3 gain), and was not detrimental to ESIM. However, in case of the $\text{fastText}_{[\text{MIMIC-III}]}$ embeddings knowledge-directed attention was beneficial to both models, as shown in Table 7. Note that while retrofitting can use only direct relations during the training process, our method incorporates information about relationships of any length, which is a necessity (as evident from Figure 3).

Embedding	InferSent	ESIM
$\text{GloVe}_{[\text{CC}]}$	0.3	0.0
$\text{fastText}_{[\text{MIMIC-III}]}$	0.2	0.3

Table 7: Absolute gain in accuracy using knowledge-directed attention.

6 Error analysis

The neutral class is the hardest to recognize for all models. Majority errors stem from confusion between entailment and the neutral class. Use of domain-specific embeddings trained on MIMIC-

Category	Premise	Hypothesis	Predicted	Expected
Numerical Reasoning	WBC 12 , Hct 41 .	WBC slightly elevated	contradiction	entailment
World Knowlegde	No known sick contacts	No recent travel	entailment	neutral
Abbreviations	No CP or fevers.	Patient has no angina	neutral	entailment
Medical Knowledge	EKG showed T-wave depression in V3-5, with no prior EKG for comparison.	Patient has a normal EKG	neutral	contradiction
Negations	Head CT was negative for bleed.	The patient has intracranial hemorrhage	neutral	contradiction

Table 8: Representative errors made by different models

III result in gains which are equally distributed across all three classes. Interestingly, gains from knowledge-directed attention stem mostly (60%) from the neutral class. Moreover, 87% of these neutral predictions were predicted as entailment before adding the knowledge directed attention.

We categorized the errors made by all the models in four broad categories. Table 8 outlines representative errors made by most models in these categories. Numerical reasoning such as *abnormal lab value* $\rightarrow$ *disease* or *abnormal vital sign* $\rightarrow$ *finding* are very hard for a model to learn unless it has seen multiple instances of the same numerical value.[7] The first step is to learn what values are *abnormal* and the next is to actually perform the inference. Many inferences require world knowledge that could be deemed close to open domain NLI. While these are very subtle, some are quite domain specific (e.g. *emergency admission* $\nrightarrow$ *planned visit*). Abbreviations are ubiquitously found in clinical text. While some are standard and therefore frequent, clinicians tend to use non standard abbreviations making inference harder. Finally, many inferences are at the core of reasoning with clinical knowledge. While training on large datasets maybe a natural but impractical solution, this is an open research problem for researchers in the community.

7 Limitations

Unlike SNLI and MultiNLI, each example in the MedNLI dataset was single annotated. However, this was the best we could do in the limited time and resources available. Very recently Gururangan et al. (2018) discovered annotation artifacts in NLI datasets. Since we followed the exact same process, we found them to be present in MedNLI as well. The premise-oblivious text-classifier that achieves 67.0 F1 on SNLI, and 53.9 on Multi-NLI achieves 61.9 on MedNLI.

[7]The symbol $\rightarrow$ represents entailment relationship

8 Conclusion

We have presented MedNLI, an expert annotated, public dataset for natural language inference in the clinical domain. To the best of our knowledge, MedNLI is the first dataset of its kind. Our experiments with several state-of-the-art models provide a strong baseline for this dataset. Our work compliments the current efforts in NLI by presenting thorough experiments for the specialized and knowledge intensive field of medicine. We also demonstrated that a simple use of domain-specific word embeddings provides a performance boost. Finally, we also presented a method for integrating domain ontologies into the training regime of models. We hope the released code and dataset with clear benchmarks help advance research in clinical NLP and the NLI task.

Acknowledgments

This work would not have been possible without Adam Coy, Andrew Colucci, Chanida Thammachart, and Hassan Ahmad – the clinicians in our team who helped us in creating the dataset. We are grateful to Vandana Mukherjee and Tanveer Syeda-Mahmood for supporting the project. We would also like to thank Anna Rumshisky and Anna Rogers for their help in this work. Most importantly, we would like to thank Leo Anthony Celi and Alistair Johnson from the MIMIC team for helping us in making MedNLI publicly available.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained anal-

ysis of sentence embeddings using auxiliary prediction tasks. In *Proceedings of ICLR*.

Alan R Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR*.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*.

Anita Burgun and Olivier Bodenreider. 2001. Comparing terms, concepts and semantic classes in WordNet and the Unified Medical Language System. In *Proceedings of the NAACL Workshop: WordNet and other lexical resources: Applications, extensions and customizations.*, pages 77–82.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced lstm for natural language inference. In *Procedings of ACL*.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of EMNLP*.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 15(4):i–xvii.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer.

Manaal Faruqui, Jesse Dodge, Sujay K. Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL*.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Carol Friedman, Pauline Kra, and Andrey Rzhetsky. 2002. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, 35(4):222–235.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 315–323.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. *Proceedings of NAACL*.

Paul KJ Han, William MP Klein, and Neeraj K Arora. 2011. Varieties of uncertainty in health care: a conceptual taxonomy. *Medical Decision Making*, 31(6):828–838.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. 2018. Transfer learning for named-entity recognition with neural networks. *Proceedings of LREC*.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

Lingpipe. 2008. LingPipe 4.1.0.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Proceedings of NIPS*.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Stephane M. Meystre, Guergana K. Savova, Karin C. Kipper-Schuler, and John F. Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of Medical Informatics*, 35:128–44.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of EMNLP*.

Ted Pedersen, Serguei VS Pakhomov, Siddharth Patwardhan, and Christopher G Chute. 2007. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3):288–299.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2016. Reasoning about entailment with neural attention. In *Proceedings of ICLR*.

Sunil Kumar Sahu and Ashish Anand. 2017. What matters in a transferable neural network model for relation classification in the biomedical domain? *arXiv preprint arXiv:1708.03446*.

Chaitanya Shivade, Courtney Hebert, Marcelo Lopetegui, Marie-Catherine De Marneffe, Eric Fosler-Lussier, and Albert M Lai. 2015. Textual inference for eligibility criteria resolution in clinical trials. *Journal of Biomedical Informatics*, 58:S211–S218.

Gabriel Stanovsky, Daniel Gruhl, and Pablo Mendes. 2017. Recognizing mentions of adverse drug reaction in social media using knowledge-infused recurrent models. In *Proceedings of EACL*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1):138.

Oriol Vinyals, Łukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. 2015. Grammar as a foreign language. In *Advances in Neural Information Processing Systems*, pages 2773–2781.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of EMNLP*.