Clinical Contradiction Detection

Anonymous ACL submission

Abstract

Detecting contradictions in text is essential in determining the validity of the literature and sources that we consume. Medical corpora are riddled with conflicting statements. This is due to the large throughput of new studies and the difficulty in replicating experiments, such as clinical trials. Detecting contradictions in this domain is hard since it requires clinical expertise. In this work, we present a distant supervision approach that leverages a medical ontology to build a seed of potential clinical contradictions over 22 million medical abstracts. As a result, we automatically build a labeled training dataset consisting of paired clinical sentences that are grounded in an ontology and represent potential medical contradiction. The dataset is used to weakly-supervise state-of-the-art deep learning models showing significant empirical improvements across multiple medical contradiction datasets.

1 Introduction

Determining whether a pair of statements is contradictory is foundational to fields including science, politics, and economics. Detecting that statements contradict can shed light on fundamental issues. For instance, mammography is an integral routine in modern cancer risk detection, but there is conflicting material about its efficacy (Boyd et al., 1984). Recognizing that a certain topic has opposing points of view, signifies that this issue may deserve further investigation. Medicine is a particularly interesting domain for contradiction detection, as it is rapidly developing, of high impact, and requires an above-superficial understanding of the text. According to the National Library of Medicine, the PubMed (Canese and Weis, 2013) database averaged 900k citations for the years 2018-2021, with a quickly growing trajectory (med, 2006). The publication of contradictory papers is not uncommon in scientific research, as it is part of the process of validating or refuting hypotheses and advancing knowledge in a field. A study on highly impactful clinical research found that that 16% of established interventions were refuted (Ioannidis, 2005). Extrapolating these statistics to PubMed, over 5 million articles would disagree with a previous finding.

The problem of contradiction detection in text has been studied in the task of natural language inference (NLI). This task was developed to tackle the problem of recognizing whether a pair of sentences are contradictory, entailing, or neutral in text. Deep learning approaches have reached impressive results for this task. Specifically, large models with hundreds of millions of parameters such as DeBERTa (He et al., 2020) and BioELECTRA (raj Kanakarajan et al., 2021), are considered today the state-of-the-art (SOTA) for this task. However, in clinical text, defining and detecting a contradiction is more difficult. Sometimes more context may be needed in order to detect contradiction due to the high difficulty of the material. Consider the following example:

1. “However, in the valsartan group, significant **improvements in left ventricular hypertrophy and microalbuminuria** were observed.”
2. “Although a bedtime dose of doxazosin can significantly **lower the blood pressure**, it can also **increase left ventricular diameter**, thus **increasing the risk of congestive heart failure.**”

Detecting that this pair contradicts requires knowing that *improvements in left ventricular hypertrophy* is a positive outcome, whereas an *increase [in] left ventricular diameter* is negative outcome with regards to heart failure.

To tackle contradiction detection using deep learning methods, large contradiction datasets are required. However, very few datasets exist to train such algorithms in the clinical contradiction domain. One reason for this could be due to the time and cost of labeling complex medical corpora. The MedNLI dataset (Romanov and Shivade, 2018) for instance, required the expert labeling of 4 clinicians over the course of 6 weeks [1]. Yet, MedNLI is fabricated in the sense that each of the clinicians was given a clinical description of a patient and came up with a contradicting, entailing, and neutral sentence to pair up with that description. However, in this work we are more interested in naturally-occurring sentences in clinical literature as opposed to manually curated texts that will not be representative. Specifically, we focus on sentences representing clinical outcomes and attempt to identify whether they are contradictory.

One of the approaches to overcome the lack of large enough data is distant supervision (Mintz et al., 2009). Distant supervision is a technique for training machine learning models on a large corpus of data without manual annotation. It works by using existing knowledge sources (such as a database of facts) to automatically label a large amount of data. The quality of the labels can be noisy, so the goal is to train models that are robust and can still learn meaningful patterns. We propose a novel methodology leveraging distant supervision and a clinical ontology - the Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT or SNOMED for short) (Stearns et al., 2001). SNOMED is developed by a large and diverse group of medical experts (Donnelly et al., 2006) and it contains extensive information about clinical terms and their relationships. Our methodology uses knowledge extracted from SNOMED to classify pairs of "naturally-occurring", potentially contradictory sentences. PubMed's database of medical abstracts is our source for naturally-occurring sentences.

We perform empirical evaluation over multiple manually labeled clinical contradiction datasets. We fine tune SOTA deep learning models on the aforementioned ontology-driven created dataset. The results demonstrate that the distant-supervision-based methodology we propose yields statistically significant improvements of the models for contradiction detection. The average results of 8 different models see an improvement on our main evaluation set (Section 4.1.1) over previous SOTA. Specifically, we find that the improvement is consistent across both small models and those that are considered to be SOTA on NLI tasks, which is the closest task to that of contradiction detection.

The contribution of our work is threefold: (1) We present the novel problem of contradiction analysis of naturally-occurring sentences in clinical data. (2) We create a clinical contradiction dataset through the use of distant supervision over a clinical ontology which yields improvements of SOTA deep learning models when fine-tuning on it. (3) We perform empirical evaluation over numerous manually labeled clincal contradiction datasets showing improvements of SOTA models when fine-tuned on the ontology-driven dataset.

2 Related Work

The field of natural language inference has primarily focused on textual entailment with the RTE challenges proposed by Dagan et al. (2013) and Dagan et al. (2005). The task involves determining if the meaning of one sentence can be inferred from another. Over time, new data and classification criteria have been introduced, including the labeling of contradictions in the third challenge (Giampiccolo et al., 2007). However, the medical domain brings additional challenges for contradiction detection requiring clinical expertise.

Despite the complexity of the medical literature and the reality of contradictions amongst publications, there has been surprisingly little work in this area. Large NLI corpora contain relatively easy contradiction pairs, partly due to the cost of annotating complex contradictions. The contradiction is often a negation through words like 'not'. An example from a large NLI corpus, MultiNLI (Williams et al., 2017) is:

1. "**Met my** first girlfriend that way."
2. "**I didn't meet my** first girlfriend until later."

Alamri and Stevenson (2016) developed a dataset labeled for contradictory research claims in abstracts related to cardiovascular medicine. This corpus has more complex sentence-pairings and is annotated by experts in the field.

Some works addressed contradiction of a clinical query and a claim. Given a sentence and a question, Tawfik and Spruit (2018) use a combination of hand-crafted features to build a classifier, whereas (Yazi et al., 2021) use pure deep neural network (DNN) techniques. Unlike these approaches, we focus on classifying any given pair of medical sentences representing a clinical outcome. To the best of our knowledge no work addresses contradiction detection between naturally-occurring sentences in

[1] To access MedNLI, users must be MIMIC-III certified.

clinical literature.

This works leverages distant supervision (Mintz et al., 2009) to address the task of identifying contradiction detection between clinical sentence-pairs representing clinical outcomes. We propose to weakly-supervise SOTA deep learning models during fine-tuning by utilizing the relational knowledge of a clinical ontology. Unlike common distant supervision approaches (Smirnova and Cudré-Mauroux, 2018; Purver and Battersby, 2012), we do not use a database with known relationship labels, but instead use the structure and attributes of the clinical ontology to infer whether terms are contradictory. To the best of our knowledge, our work is the first time distant supervision is used for contradiction detection in the clinical realm.

3 Methods

We aim to create a model for accurately classifying whether two clinical outcomes contradict. In particular, we focus on examples which are non-trivial and require a deeper understanding of the subject area or text. This model brings awareness to conflicting literature and findings, specifically in the medical domain. Understanding where there is disagreement, can help elicit further investigations or general consciousness.

3.1 SNOMED CT Ontology

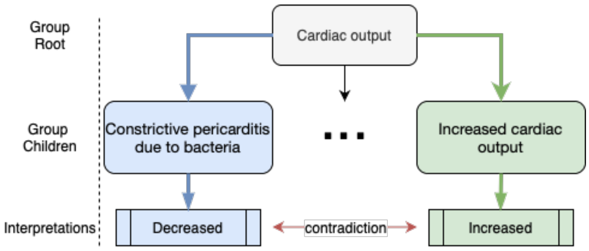

Figure 1: The group with *Cardiac output* as its root. The children depicted have contradicting interpretations.

SNOMED is an ontology containing over 350,000 clinical terms (Stearns et al., 2001). The terminology contains information about a plethora of health concepts, often containing useful attributes such as relationships to other terms and various interpretations. The structure of SNOMED allows us to group terms based on their relationships. We hypothesize that that using this structure coupled with synonyms and antonyms, will enable us to create a corpora of contradicting and non-contradicting clinical terms.

3.1.1 SNOMED Node Attributes

Each term in the SNOMED ontology is a node in a tree-like structure. A subset of these nodes have useful attributes which we use to determine their inter-relationships. Each of these nodes belongs to a group which is parented by the group root. In addition, each node has a simple interpretation. In Figure 1, the group consists of nodes describing the group root *cardiac output*. The green (left) node, *increased cardiac output*, has the interpretation - *increased*.

We claim that every grouping of terms which has these attributes has a logical connection. We argue that pairing up child nodes yields a natural combination of contradicting and non-contradicting pairs of phrases. Determining the relationship between a pair of SNOMED terms is done partially through comparing their interpretations. In Figure 1 the blue (left) node has the interpretation *decreased*, whereas the green (right) node has the interpretation *increased*. Since the values of these fields are different, we assign the pair an *attribute* label ($A_{i,j}$) of contradiction. In Algorithm 1, $A_{i,j}$ is assigned on Line 12.

The size of the groupings can get large. For instance, the group root *Cardiac function* has 275 children. Since *cardiac function* is very general, its child terms may not be related - for example the terms *aortic valve regurgitation due to dissection* and *dynamic subaortic stenosis*. Both terms are impairments of *cardiac function*, but it would not be fair to claim that the two are related outcomes. Though these large groupings can yield many pairings of phrases, we see why they may also be less accurate. Some of this testing is in Section 5.2, where we investigate the effects of group sizes.

Below we include pairings of contradictions in various medical domains that our methodology yields:

- suppressed urine secretion ↔ polyuria
- elevation of SaO2 ↔ oxygen saturation within reference range
- joint stable ↔ chronic instability of joint

3.1.2 Synonyms

After exploiting ontological structure, we consider linguistic elements. Although synonyms and antonyms do not always indicate whether sequences of words are contradictory, they provide a strong signal in our structural construction. Since clinical terms are already grouped, we know that all the terms in a grouping share a context, thereby

Algorithm 1 SNOMED Traversal

```
1:  function TRAVERSE($root$)
2:      for $n \in root.children$ do
3:          if $n.num\_childs \leq group\_size$
4:              $pairs \leftarrow$ DET_RELATION($n$)
5:          end if
6:      end for
7:      return $pairs$
8:  end function

9:  function DET_RELATION($n$)
10:     $pairs \leftarrow \{\}$
11:     for $c_i, c_j \in n.child\_pairs$ do
12:         $A_{i,j} \leftarrow$ GET_ATTR_LABEL($c_i, c_j$)
13:         $S_{i,j} \leftarrow$ GET_SYN_LABEL($c_i, c_j$)
14:         $label_{i,j} \leftarrow A_{i,j}$
15:         if $S_{i,j}$ = no-contra &$A_{i,j}$ = contra
16:             $label_{i,j} \leftarrow$ contra
17:         else if $S_{i,j}$ = contra &$A_{i,j}$ = no-contra
18:             $label_{i,j} \leftarrow$ contra
19:         end if
20:         $pairs \leftarrow pairs \cup \{(label_{i,j}, c_i, c_j)\}$
21:     end for
22:     return $pairs$
23: end function

24: $SNOMED \leftarrow$ TRAVERSE($root$)
25: FINETUNE($Model, SNOMED$)
```

allowing the use of simpler indicators to determine their relationship. We word-tokenize each clinical phrase, removing the intersection of the two sets of tokens, leaving each set with its unique tokens.

Figure 2 illustrates this for the clinical terms *shortened p wave* and *prolonged p wave*. The respective unique tokens are *shortened* and *prolonged*. Since the unique tokens are antonyms, the *synonym* label for the pair is a contradiction. In Algorithm 1, the *synonym* label ($S_{i,j}$) is assigned on Line 13. Similarly, if the respective tokens are synonyms, then $S_{i,j}$ would be a non-contradiction.

3.1.3 Combining Attributes and Synonyms

To optimally combine $A_{i,j}$ and $S_{i,j}$ to form a final $label_{i,j}$, we build a validation set of the publicly available SNOMED term-pairs. A human annotator with an advanced medical degree (M.D.) labeled 149 SNOMED phrase-pairs - 70 of which were contradictory and 79 as non-contradictory. More details can be found in Appendix A.1. We find that when $A_{i,j}$ indicates contradiction, then it's highly likely that $label_{i,j}$ is a contradiction. The same is true if $S_{i,j}$ indicates contradiction. We define the explicit logic in Lines 15 through 19. We reach 79% accuracy through using heuristics on the human-labeled SNOMED term-pairs in the validation dataset.

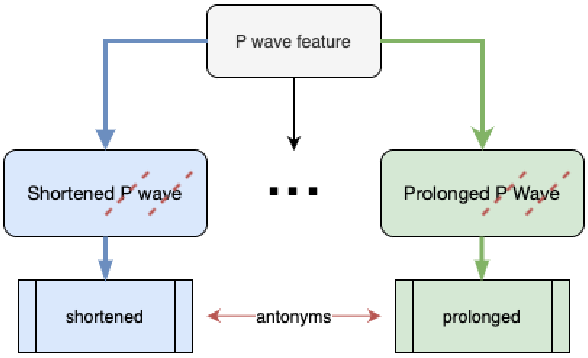

Figure 2: The terms *shortened p wave* and *prolonged p wave* are simplified to *shortened* and *prolonged*. The remaining words are antonyms.

3.2 Ontology-Driven Distant Supervision

Using the relational knowledge extracted from the SNOMED ontology, we weakly-supervise naturally-occurring sentences in PubMed to build our SNOMED dataset. We fine-tune on this dataset to achieve significant improvements over existing baselines. Algorithm 1 summarizes the procedure. We search PubMed for sentences containing the phrase-pairs discussed in Section 3.1, resulting in a corpus of pairs of sentences. The sentence-pairs are then labeled through distant supervision as explained below. For a given pair of SNOMED terms (p_1, p_2), we label sentences (s_1, s_2) as formalized in Eq.1, where $label \in$ {contradiction, non-contradiction}.

$$(p_1 \in s_1) \wedge (p_2 \in s2) \wedge ((p_1, p_2) \in label) \quad (1)$$

Naively, we pair-up any sentences satisfying Equation 1, independent of whether they appear in the same text, when creating the SNOMED dataset. Although two sentences contain their respective clinical SNOMED terms, they may be unrelated. The sentence-pair below exhibits this:

1. “The present results suggest that the upstream changes in blood flow are transmitted by the velocity **pulse faster** than by the pressure pulse in the microvasculature.”
2. “His chest wall was tender and his **pulse slow** but the remainder of his physical examination was normal.”

The bolded clinical terms are central to the meaning of the sentences and are independently contradictory. However, when placed in context they may be less relevant to each other as in the example above. We experiment through imposing stricter criteria for filtering sentence matches - namely MeSH (Medical Subject Headings) terms criteria (Lipscomb, 2000) and cosine similarity criteria.

MeSH terms are words used to categorize articles within PubMed. We hypothesize that if sentences are drawn from articles with related MeSH terms, then the likelihood that they discuss the same topic increases. Equation 2 is the formulation we use for filtering via MeSH terms. $MeSH_i$ and $MeSH_j$ are the sets of MeSH terms for articles containing $sent_i$ and $sent_j$ respectively. Let t be a chosen threshold.

$$\mathbf{1}_A := \begin{cases} 1 & \text{if } \frac{|MeSH_i \cup MeSH_j|}{min(|MeSH_i|,|MeSH_j|)} \geq t \ , \\ 0 & otherwise \end{cases} \tag{2}$$

Although MeSH terms are powerful, they are not perfect. The following sentence-pair achieves a score of 0.4 per the inequality in Equation 2.

1. "In dogs challenged with endotoxin, the inhibition of nitric oxide production **decreased cardiac index** and did not improve survival."
2. "Intra-aortic balloon pumping **increased cardiac index** and aortic distensibility by 24% and 30%, respectively, and reduced myocardial oxygen demand by 31% ($P < .001$ for all alterations)."

Despite overlap in MeSH terms, they are very different - one discusses dogs and the other humans.

The second filtration method measures the cosine similarity between one-hot vectors. Topically related sentences should have a higher one-hot vector cosine similarity. Let $onehot_i$ and $onehot_j$ be the respective one-hot vectors of $sent_i$ and $sent_j$. Note vectors lengths are equal to the number of unique words spanning the sentence-pair. We compute the cosine similarities between the vectors as shown in equation 3. For the dog example above, cosine similarity yields a score of 0.2.

$$\mathbf{1}_A := \begin{cases} 1 & \text{if } cossim(onehot_i, onehot_j) \geq t \ , \\ 0 & otherwise \end{cases} \tag{3}$$

4 Empirical Evaluation

In this section we discuss the medical corpora used in our evaluation of 8 different models, spanning various model sizes and objective functions.

Table 1: Cardiology Dataset Breakdown

Split	Total	Contra	Non-Contra
Train	1347	571	776
Dev	198	100	98
Test	227	55	172

4.1 Evaluation Datasets

In determining whether our methodology can provide reliable results, we acquire and modify various medically related corpora.

4.1.1 Cardiology Dataset

Due to the difficulty of labeling medical data, there are few datasets labeled for medical contradictions. To evaluate results and compare the quality of the dataset we create in an automatic fashion, we tweak an existing cardiology dataset. Alamri et al. developed ManConCorpus (Alamri and Stevenson, 2016) - a dataset of potentially contradictory research claims in abstracts related to cardiovascular medicine. The corpus is composed of question-claims pairs. Each question has multiple 'yes', 'no' claims. The claims are naturally-occurring sentences in PubMed, whereas the questions are generated by expert labelers. We convert ManConCorpus by pairing up the claims, since we are strictly interested in naturally-occurring sentences from PubMed. A pair is labeled as contradictory if each constituent claim answers the question differently. We coin this dataset as the Cardiology Dataset ("Cardio") (see Table 1 for details).

4.1.2 Hard Cardiology Dataset

Through our analysis, we find that models tend to classify sentence-pairs as contradictory if negation words appear. For example:

1. "Our results indicate that atorvastatin therapy significantly improves BP control in hyperlipidemic hypertensive patients."
2. "Administration of a statin in hypertensive patients in whom blood pressure is effectively reduced by concomitant antihypertensive treatment **does not have** an additional blood pressure lowering effect."

Thus, we construct a version of Cardio through rewriting the sentences without negation words. As expected, this version exposes some of the weaknesses of the models, since negation words are no longer deemed as important.

4.1.3 MedNLI

Inspired by SNLI (Bowman et al., 2015), MedNLI was created similarly, but with a focus on the clinical domain (Romanov and Shivade, 2018). The dataset was curated over the course of six weeks, borrowing the time of four doctors. MedNLI consists of sentence-pairs which are grouped into triples - a contradictory, entailing, and neutral pair. The sentences are not naturally-occurring in existing medical literature. The premise is shared across the three pairs, but each have a different hypothesis, yielding a different label. Since MedNLI deals with a 3-class problem, we relabel the dataset by making $\{$*entailment, neutral*$\}$ map to *non-contradiction*.

Our focus is to show that the SNOMED dataset, which requires no expert intervention or expenses, is as powerful as the curated MedNLI dataset. We find that the baseline on the relabeled version of MedNLI gives high results (0.974), so adding additional data makes little change. The largest labeled datasets containing naturally-occurring sentences are at most hundreds of sentences. Therefore, we randomly sample 100 instances from MedNLI's train-split and report results on that.

To explore fields outside of cardiology, we create versions of MedNLI focused on gynecology (GN), endocrinology (Endo), obstetrics (OB), and surgery. To filter the data, we use the help of the same annotator introduced in Section 3.1.3. We sample from the train-split in the same fashion as explained above. Note that these datasets also have the same 2-class label structure as explained in Section 4.1.3. More details are found in Appendix A.2.

4.2 Baseline Models

Yazi et al. (2021) achieve the SOTA on the ManConCorpus, which we turn into the Cardio corpus as explained in Section 4.1.1. They concatenate BERT embeddings for their question and claim, feeding this input into a multi-layer feed forward network. All of our baselines do not use a siamese network, instead we feed in our sentence-pairs as input into the network. Our evaluation consists of 8 baseline models and comparing their performance when they are fine-tuned on the SNOMED dataset versus without. The task of classifying contradiction is most similar to NLI, so some of these baseline models are those that top leaderboards for the MNLI and MedNLI datasets - namely DeBERTaV3-Base (He et al., 2021), ALBERT (Lan et al., 2019), and BioELECTRA (raj Kanakarajan et al., 2021). ELECTRA (Clark et al., 2020) and BERT (Devlin et al., 2018) are also included as they are generally high-performing architectures. In addition, we are interested in seeing the performance of small models. They require less computing resources and may allow the SNOMED dataset to have a stronger influence during fine-tuning. Thus, we also include BERT-Small (Turc et al., 2019), ELECTRA-Small, and DeBERTaV3-Small (He et al., 2021). Table 2 contains a breakdown of the number of parameters per model.

Table 2: Baseline Models Parameter Count

Model	Parameter Count
ALBERT	11.7M
ELECTRA-Small	13.5M
BERT-Small	28.8M
ELECTRA-Base	109.5M
BERT-Base	109.5M
BioELECTRA	109.5M
DeBERTaV3-Small	141.9M
DeBERTaV3-Base	184.4M

All the baseline models are pre-trained on large corpora. The high-level architecture of the models is the same, so we use the functionalities of HuggingFace (Wolf et al., 2019) and the Sentence-Transformer library (Reimers and Gurevych, 2019). We add an uninitialized binary classification head on top of the model body. We adopt all the hyperparameters from the Sentence-Transformer library with the exception of training batch size, which is set to 8 for models above 30M parameters and to 16 for models under 30M parameters.

Each baseline we tune with the SNOMED dataset. The SNOMED dataset we create uses a group size of 25, sampling 10 sentence-pairs from PubMed for every SNOMED term-pair. These hyperparameters are determined through ablation tests on the Cardio validation set.

5 Empirical Results

The following sections display the significance of the SNOMED dataset we create via the methodology (Section 3). We explore additional insights through ablation tests and qualitative examples.

5.1 Main Result

Table 3 summarizes our main findings. We compare the performance of the baseline algorithms when fine-tuned over the original training of each dataset (marked as "base") versus tuning using our

Table 3: Performance of Models tuned with SNOMED vs. Without

		Algorithm								
Dataset	Method	ALBERT Base	ELECTRA Small	BERT Small	ELECTRA Base	BERT Base	Bio-ELECTRA	DeBERTa Small	DeBERTa Base	(Yazi et al., 2021)
Cardio	Base	0.911	0.877	0.858	0.863	**0.914**	0.880	0.885	0.861	0.858
	Ours	**0.928**	**0.947***	**0.958***	**0.892**	0.878	**0.925**	**0.931***	**0.942***	-
Hard-Cardio	Base	0.876	0.785	0.717	0.847	**0.803**	0.850	0.842	0.845	0.687
	Ours	**0.925***	**0.853***	**0.794***	**0.873**	0.791	**0.925***	**0.917***	**0.936***	-
MedNLI-General	Base	0.598	0.526	0.537	0.587	0.651	0.616	0.585	0.696	0.528
	Ours	**0.780***	**0.615***	**0.656***	**0.789***	**0.764***	**0.798***	**0.778***	**0.876***	-
MedNLI-Cardio	Base	0.638	0.524	0.555	0.599	0.675	0.607	0.601	0.673	0.585
	Ours	**0.789***	**0.668***	**0.727***	**0.780***	**0.793***	**0.795***	**0.796***	**0.875***	-
MedNLI-GYN	Base	0.642	0.492	0.608	0.692	0.683	0.575	0.592	0.633	0.615
	Ours	**0.692***	**0.633***	**0.667***	**0.800***	**0.817***	**0.775***	**0.608**	**0.825***	-
MedNLI-Endo	Base	0.568	0.494	0.551	0.575	0.722	0.631	0.605	0.625	0.549
	Ours	**0.801***	**0.607***	**0.702**	**0.811***	**0.852***	**0.893***	**0.728***	**0.909***	-
MedNLI-OB	Base	0.514	0.521	0.541	0.573	0.560	0.506	0.527	0.526	0.542
	Ours	**0.657***	**0.545**	**0.557**	**0.693***	**0.644***	**0.698***	**0.590**	**0.750***	-
MedNLI-Surgery	Base	0.641	0.519	0.528	0.597	0.919	0.665	0.640	0.752	0.539
	Ours	**0.890***	**0.739***	**0.802***	**0.860***	**0.929***	**0.903***	**0.885***	**0.922***	-

novel SNOMED dataset and the training dataset (marked as "base+SNOMED"). We measure the area under the ROC curve of each baseline, and verify statistical significance through Delong's test (DeLong et al., 1988). Significant differences are marked with an asterisk (*). We observe that across all dataset the weak supervision over the SNOMED dataset reached superior results compared to fine tuning only on the original dataset and outperforms the SOTA model for contradiction detection (Yazi et al., 2021).

Cardio is a relatively difficult dataset of potentially contradicting pairs of sentences naturally-occurring in PubMed. The sentences are complex and require a deep medical understanding. We observe that fine tuning on the SNOMED dataset improves the baselines for all 7 out of the 8 models we evaluate over the Cardio dataset.

The performance on Hard-Cardio drops relatively to Cardio as expected. This verifies our hypothesis that removing negations makes the problem more difficult. Further, 7 out 8 models fine-tuned on SNOMED outperform their baseline counterparts.

We observe that even on synthetically created common datasets, such as MedNLI sentences, our methodology improves over *all* baselines for this corpus. We observe a similar trend when focusing on various sub-specialties. The improvements are consistent across *all* models when fine-tuning on SNOMED. This enables us to learn of the scalability of our methods for clinical contradiction detection through different fields within healthcare.

Analyzing our findings further, we see that there is a trend that smaller models are generally more affected by fine-tuning on SNOMED. All of the evaluation datasets improve over the baseline on *every* model under 30 million parameters.

5.2 Ablation Studies

In this section we review the ablation studies to determine potential impact of the different parameters of the system on performance.

5.2.1 Group and Sentence Samples Size

We explain SNOMED term grouping in Section 3.1 and illustrate in Figure 1. The size of a group and the quality of the pairing may be closely related. Larger groupings tend to have more terms which are less directly related to each other as explained in Section 3.1.1. Thus, we experiment with creating SNOMED datasets based on terms belonging to groups of at most 6, 12, 25, and 50 terms.

During dataset creation, we choose how many sentence-pairs to sample per SNOMED pairing. In Figure 3, each line with a different color/marker represents a different number of samples averaged across all 8 models. The ablations we perform include 10, 25, and 50 samples per pairing.

Figure 3 shows 10 samples outperforms higher sampling numbers for almost all group numbers. Increased sampling results in over-saturation of certain term-pairs. This may result in overfitting. The best group size is 25 for small models and 12 for large models. These numbers strike the balance of creating a large amount of SNOMED phrase-pairings, while keeping their relationships accurate (as discussed in Section 3.1.1). Smaller models may benefit more from larger group sizes, because they have a more limited base knowledge than those of large models.

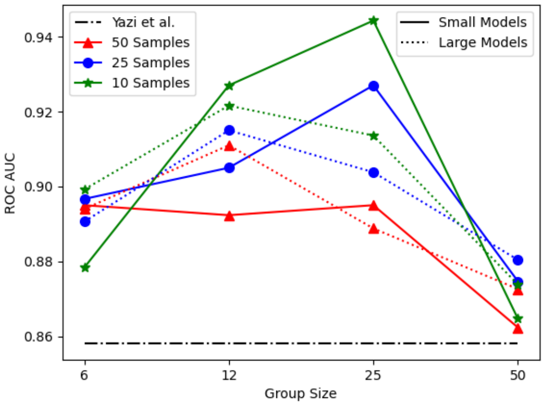

Figure 3: Small and large model performance across group sizes and sample numbers. Reported on Cardio.

5.2.2 Filtering Based on Similarity

To increase the chances that sentences are related, when sampling phrases from PubMed, we experiment with keeping pairs that exhibit high MeSH term or cosine similarity as explained in Section 3.2. Figure 4 shows the relationship between the filtration methods discussed above. As a continuation of the ablation visualized in Figure 3, we fix the number of samples to be 10 and the group size to be 25. The cosine methodology outperforms both the naive version (no filtering) and MeSH. Although MeSH terms are useful, it is possible that since they are tagged on an article-level, they cannot provide the same topic granularity as the one-hot vectors.

6 Conclusions

Contradiction detection is central to many fields, but it is especially important in medicine due to direct human impact. With the rapid growth of the field, clinical research is exploding with new findings as demonstrated by the growth of PubMed. Although contradictions are a subfield of NLI, there is much less exploration in the clinical domain. Often times, contradictions within medicine are more complex than other fields due to the need of additional context and domain knowledge. Labeling datasets which could produce high performing results with deep learning models are time and resource costly.

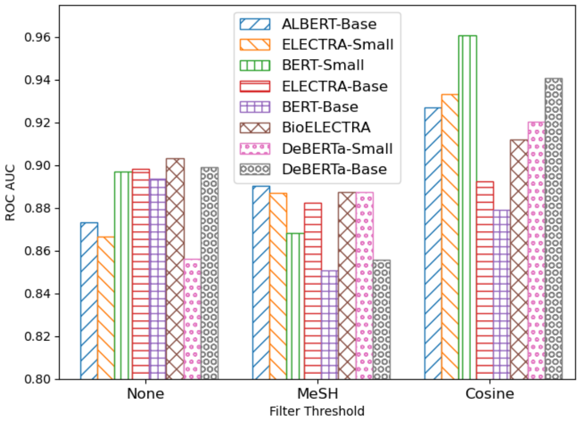

Figure 4: Model performance across varying filtration methods. Number of samples is 10 and group size is 25 for plotted results. Reported on Cardio.

We introduced a novel methodology of using a clinical ontology to weakly-supervise the creation of a contradiction dataset with naturally-occurring sentences. We coin this dataset as the SNOMED dataset. The empirical results suggest that fine-tuning on the SNOMED dataset results in consistent improvement across multiple SOTA models over diverse evaluation datasets spanning multiple medical specialties. We showed that a balance exists between the group size of the number of terms and the number of sentences sampled from PubMed per term-pairing. In addition, we find that we can further improve results through filtering which PubMed sentences we include in our dataset.

For future exploration we suggest investigating more robust sentence filtration methods, such as topic modeling or sentence embedding similarity. Looking into how other clinical ontologies can be paired with SNOMED may also be fruitful.

Limitations

The methodology proposed is limited to using clinical terms which are located within SNOMED. In addition, many SNOMED terms do not appear exactly within PubMed, so not all of the terms are used. Finally, the relationships we extract from the clinical ontology are not ground-truth, yielding noise during dataset creation.

Ethical Considerations

Whenever working within the clinical domain, ethical considerations are crucial. The data that we work with is all rooted in already publicly available corpora and PubMed. To the best of our knowledge the data we use does not contain any personal information of any humans involved in clinical trials. There is a potential risk of over representing common diseases and outcomes in our dataset, thereby not including enough data about other outcomes.

References

2006. Medline® citation counts by year of publication (as of january 2022)*.

Abdulaziz Alamri and Mark Stevenson. 2016. A corpus of potentially contradictory research claims from cardiovascular research abstracts. *Journal of biomedical semantics*, 7(1):1–9.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Norman F Boyd, Brian O'Sullivan, Eve Fishell, Imre Simor, and Gabriel Cooke. 1984. Mammographic patterns and breast cancer risk: methodologic standards and contradictory results. *Journal of the National Cancer Institute*, 72(6):1253–1259.

Kathi Canese and Sarah Weis. 2013. Pubmed: the bibliographic database. *The NCBI handbook*, 2(1).

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Kevin Donnelly et al. 2006. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and William B Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

John PA Ioannidis. 2005. Contradicted and initially stronger effects in highly cited clinical research. *Jama*, 294(2):218–228.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Carolyn E Lipscomb. 2000. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

Matthew Purver and Stuart Battersby. 2012. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 482–491.

Kamal raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. Bioelectra: pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. *arXiv preprint arXiv:1808.06752*.

Alisa Smirnova and Philippe Cudré-Mauroux. 2018. Relation extraction using distant supervision: A survey. *ACM Computing Surveys (CSUR)*, 51(5):1–35.

Michael Q Stearns, Colin Price, Kent A Spackman, and Amy Y Wang. 2001. Snomed clinical terms: overview of the development process and project status. In *Proceedings of the AMIA Symposium*, page 662. American Medical Informatics Association.

Noha S Tawfik and Marco R Spruit. 2018. Automated contradiction detection in biomedical literature. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 138–148. Springer.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Fatin Syafiqah Yazi, Wan-Tze Vong, Valliappan Raman, Patrick Hang Hui Then, and Mukulraj J Lunia. 2021. Towards automated detection of contradictory research claims in medical literature using deep learning approach. In *2021 Fifth International Conference on Information Retrieval and Knowledge Management (CAMP)*, pages 116–121. IEEE.

A Appendix: Annotation

As mentioned in Section 3.1.3, we work with an annotator with a degree in medicine. The annotator was recruited due to their expertise in the field.

A.1 SNOMED Term-Pairs

The annotator labeled 149 SNOMED term-pairs as either contradictory or non-contradictory. They were provided with a list of pairs, without any additional information about the ontological structure they came from. This was done in order to preserve fairness and integrity during the labeling process. The instructions were to come up with a binary label for each of the pairs.

A.2 Filtering MedNLI

The human annotator also helped with coming up with a list of sub-words which served as indicators for particular fields of medicine. For example, the sub-words *vulv* and *gyno*, are indicative of gynecology. These word lists were used to create the variations of MedNLI discussed in Section 4.1.3.