SciNLI: A Corpus for Natural Language Inference on Scientific Text

Mobashir Sadat and Cornelia Caragea
Computer Science
University of Illinois Chicago
msadat3@uic.edu cornelia@uic.edu

Abstract

Existing Natural Language Inference (NLI) datasets, while being instrumental in the advancement of Natural Language Understanding (NLU) research, are not related to scientific text. In this paper, we introduce SciNLI, a large dataset for NLI that captures the formality in scientific text and contains $107,412$ sentence pairs extracted from scholarly papers on NLP and computational linguistics. Given that the text used in scientific literature differs vastly from the text used in everyday language both in terms of vocabulary and sentence structure, our dataset is well suited to serve as a benchmark for the evaluation of scientific NLU models. Our experiments show that SciNLI is harder to classify than the existing NLI datasets. Our best performing model with XLNet achieves a Macro F1 score of only 78.18% and an accuracy of 78.23% showing that there is substantial room for improvement.

1 Introduction

Natural Language Inference (NLI) or Textual Entailment (Bowman et al., 2015) aims at recognizing the semantic relationship between a pair of sentences—whether the second sentence entails the first sentence, contradicts it, or they are semantically independent. NLI was introduced (Dagan, Glickman, and Magnini, 2005) to facilitate the evaluation of Natural Language Understanding (NLU) that significantly impacts the performance of many NLP tasks such as text summarization, question answering, and commonsense reasoning.

To date, several NLI datasets are made available (Bowman et al., 2015; Williams, Nangia, and Bowman, 2018; Marelli et al., 2014; Dagan, Glickman, and Magnini, 2005). These datasets have not only been instrumental for developing and evaluating NLI models but also have been useful in advancing many other NLP areas such as: representation learning (Conneau et al., 2017), transfer learning (Pruksachatkun et al., 2020) and multi-task learning (Liu et al., 2019a).

However, despite their usefulness, none of the existing NLI datasets is related to scientific text that is found in research articles. The vocabulary as well as the structure and formality used in sentences in scientific articles are very different from the sentences used in the everyday language. Moreover, the scientific text captured in research papers brings additional challenges and complexities not only in terms of the language and its structure but also the inferences that exist in it which are not available in the existing NLI datasets. For example, a sentence can present the reasoning behind the conclusion made in the previous sentence, while other sentences indicate a contrast or entailment with the preceding sentence. These inferences are crucial for understanding, analyzing, and reasoning over scientific work (Luukkonen, 1992; Kuhn, 2012; Hall, Jurafsky, and Manning, 2008). Therefore, ideally, the scientific language inference models should be evaluated on datasets which capture these inferences and the particularities seen only in scientific text.

To this end, we seek to enable deep learning for natural language inference over scientific text by introducing SciNLI,[1] a large dataset of $107,412$ sentence pairs extracted from scientific papers related to NLP and computational linguistics (CL) and present a comprehensive investigation into the inference types that occur frequently in scientific text. To capture the inference relations which are prevalent in scientific text but are unavailable in the existing NLI datasets, we introduce two new classes—Contrasting and Reasoning. We create SciNLI by harnessing cues in our data in the form of linking phrases between contiguous sentences, which are indicative of their semantic relations and provide a way to build a labeled dataset using distant supervision (Mintz et al., 2009). Dur-

[1] https://github.com/msadat3/SciNLI

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7399 - 7409

Class	First Sentence	Second Sentence	Linking Phrase
CONTRASTING	Essentially, that work examines how a word gains new senses, and how some senses of a word may become deprecated.	here we examine how different words compete to represent the same meaning, and how the degree of success of words in that competition changes over time.	'In contrast,'
REASONING	The Lang-8 corpus has often only one corrected sentence per learner sentence, which is not enough for evaluation.	we ensured that our evaluation corpus has multiple references.	'Thus,'
ENTAILMENT	As a complementary area of investigation, a plausible direction would be to shift the focus from the decomposition of words into morphemes, to the organization of words as complete paradigms.	instead of relying on sub-word units, identify sets of words organized into morphological paradigms (Blevins, 2016)."	'That is,'
NEUTRAL	Literature on the topic of the current study spans across many areas, including verb classification, semiotics, sign language and learning.	abstract words can be more challenging to learn and memorise.	N/A

Table 1: Examples of sentence pairs from our dataset and the linking phrases used to extract them, corresponding to all four classes considered. The second sentence of each pair is shown after removing the linking phrase.

ing training, we directly utilize these (potentially noisy) sentence pairs, but to ensure a realistic evaluation of the NLI models over scientific text, we manually annotate 6,000 sentence pairs. These clean pairs are used in two splits, 2,000 pairs for development and hyper-parameter tuning and 4,000 pairs for testing. Table 1 shows examples from our dataset corresponding to all of our four classes.

We evaluate SCINLI by experimenting with traditional machine learning models using lexical and syntactic features, neural network models—BiLSTM, CBOW, CNN, and pre-trained language models—BERT (Devlin et al., 2019), SciBERT (Beltagy, Lo, and Cohan, 2019), RoBERTa (Liu et al., 2019b), and XLNet (Yang et al., 2019). Our findings suggest that: (1) SCINLI is harder to classify than other datasets for NLI; (2) Lexical features are not enough for a model to achieve satisfactory performance on SCINLI and deep semantic understanding is necessary; (3) SCINLI is well suited for evaluating scientific NLI models; and (4) Our best performing model based on XLNet shows 78.18% Macro F1 and 78.23% accuracy illustrating that SCINLI is a challenging new benchmark.

2 Related Work

To date, several datasets exist for NLI of varying size, number of labels, and degree of difficulty. Dagan, Glickman, and Magnini (2006) introduced the RTE (Recognizing Textual Entailment) dataset of text-hypothesis pairs from the general news domain and considered two labels: entailment or no-entailment (i.e., a hypothesis is true or false given a text). The RTE dataset is paramount in developing and advancing the entailment task. The SICK (Sentences Involving Compositional Knowledge) dataset introduced by Marelli et al. (2014) was created from two existing datasets of image captions and video descriptions. SICK consists of sentence pairs (premise-hypothesis) labeled as: entailment, contradiction, or neutral. Despite being instrumental in the progress of NLI, both RTE and SICK datasets are less suitable for deep learning models due to their small size.

In recent years, SNLI (Bowman et al., 2015) and MNLI (Williams, Nangia, and Bowman, 2018) are the most popular datasets for training and evaluating NLI models, in part due to their large size. Similar to SICK, SNLI is derived from an image caption dataset where the captions are used as premises and hypotheses are created by crowdworkers, with each sample being labeled as: entailment, contradiction, or neutral. MNLI is created in a similar fashion to SNLI except that the premises are extracted from sources such as face-to-face conversations, travel guides, and the 9/11 event, to make the task more challenging and suitable for domain adaptation. More recently, Nie et al. (2020) released ANLI which was created in an iterative adversarial manner where human annotators were used as adversaries to provide sentence pairs for which the state-of-the-art models make incorrect predictions. Unlike the datasets specific to classifying the relationships between two sentences, Zellers et al. (2018) combined NLI with commonsense reasoning to introduce a new task of predicting the most likely next sentence from a number of options along with their new dataset

called SWAG which was also created with an adversarial approach. However, different from ANLI, the SWAG approach was automatic. All these datasets have been widely used for evaluating NLU models and many of them appear in different NLU benchmarks such as GLUE (Wang et al., 2018) and SUPERGLUE (Wang et al., 2019).

Heretofore, Khot, Sabharwal, and Clark (2018) created the only NLI dataset related to science. Their dataset, SCITAIL was derived from a school level science question-answer corpus. As a result, the text used in SCITAIL is very different from the type of text used in scientific papers. Furthermore, the sentence pairs in SCITAIL are classified into one of two classes: entailment or no-entailment. Thus, SCITAIL does not cover all the inference relationships necessary to understand scientific text.

In other lines of research, discourse cues, e.g., linking phrases have been previously used to extract inter-sentence and/or inter-clause semantic relations in discourse parsing (Hobbs, 1978; Webber et al., 1999; Prasad et al., 2008; Jernite, Bowman, and Sontag, 2017; Nie, Bennett, and Goodman, 2019), causal inference (Do, Chan, and Roth, 2011; Radinsky, Davidovich, and Markovitch, 2012; Li et al., 2020; Dunietz, Levin, and Carbonell, 2017) and why-QA (Oh et al., 2013). However, none of the aforementioned bodies of research investigates these relations in scientific text, nor do they exploit the discourse cues to create NLI datasets. Furthermore, discourse parsing studies a broader range of semantic relations, many of which are unrelated to the task of NLI while causal inference and why-QA are limited to only cause-effect relations. In contrast to these tasks, we focus on the semantic relations which are either relevant to the task of NLI or highly frequent in scientific text and leverage linking phrases to create the first ever scientific NLI dataset, which we call SCINLI.

3 SCINLI: A New Corpus for NLI

In order to better understand the inter-sentence relationships that exist in scientific text, we started the process of creating our dataset by perusing through scientific literature with the intent of finding clues that are revealing of those relationships. We found that to have a coherent structure, authors often use different linking phrases in the beginning of sentences, which is indicative of the relationship with the preceding sentence. For example, to elaborate or make something specific, authors use linking phrases such as "In other words" or "In particular," which indicate that the sentence supports or entails the previous sentence. We also found that some linking phrases are used to indicate additional relationships that are prevalent in scientific text but are not captured in the existing NLI datasets. For instance, when a sentence starts with "Therefore" or "Thus," it indicates that the sentence is presenting a conclusion to the reasoning in the previous sentence. Similarly, the phrase "In contrast" is used to indicate that the sentence is contrasting what was said in the previous sentence.

Therefore, inspired by the framework of discourse coherence theory (Hobbs, 1978; Webber et al., 1999; Prasad et al., 2008) that characterizes the inferences between discourse units, we extend the NLI relations commonly used in prior NLI work—entailment, contradiction, and semantic independence—to a set of inference relations that manifest in scientific text—contrasting, reasoning, entailment, and semantic independence (§3.1). In order to create a large training set with minimal manual effort, we employ a distant supervision method based on linking phrases that are commonly used in scientific writing and are indicative of the semantic relationship between adjacent sentences (§3.2). We avoid the noise incurred by the distant supervision method in our development and test sets by manually annotating these sets (§3.3).

3.1 Inference Classes

We define the inference classes used to create our dataset in this section.

3.1.1 CONTRASTING

Our CONTRASTING class is an extension of the CONTRADICTION class in the existing NLI datasets. With this class, in addition to contradicting relations between sentences in a pair, we aim to capture inferences that occur when one sentence mentions a comparison, criticism, juxtaposition, or a limitation of something said in the other sentence. We can see an example of a sentence pair from our CONTRASTING class in Table 1. Here, the authors discuss how their work differs from the other work mentioned in the first sentence thereby making a comparison between the two works.

3.1.2 REASONING

The examples where the first sentence presents the reason, cause, or condition for the result or conclusion made in the second sentence are placed in

Label	Linking Phrases
CONTRASTING	'However', 'On the other hand', 'In contrast', 'On the contrary'
REASONING	'Therefore', 'Thus', 'Consequently', 'As a result', 'As a consequence', 'From here, we can infer'
ENTAILMENT	'Specifically', 'Precisely', 'In particular', 'Particularly', 'That is', 'In other words'

Table 2: Linking phrases used to extract sentence pairs and their corresponding classes.

our REASONING class. In Table 1, we can see an example where the authors mention that they use a multi-reference corpus for evaluation in the second sentence and provide the reason behind it in the first sentence.

3.1.3 ENTAILMENT

Our ENTAILMENT class includes the sentence pairs where one sentence generalizes, specifies or has an equivalent meaning with the other sentence. An example from this class can be seen in Table 1. In the example, the second sentence is specifying the proposed direction mentioned in the first sentence making the pair suitable for our ENTAILMENT class.

3.1.4 NEUTRAL

The NEUTRAL class includes the sentence pairs which are semantically independent. We can see an example from this class in Table 1. Here, the first sentence discusses the span of the literature of a particular topic, whereas the second sentence mentions the challenges of handling abstract words in certain tasks. Therefore, the sentences are semantically independent of each other.

3.2 Training Set Creation

We construct our training set from scientific papers on NLP and computational linguistics available in the ACL Anthology, published between 2000 and 2019 (Bird et al., 2008; Radev, Muthukrishnan, and Qazvinian, 2009). For extracting textual data from the PDF papers, we use GROBID[2] which is a popular tool for parsing PDF files. We employ the following distant supervision technique on the extracted text to select and label the sentence pairs.

We create a list of linking phrases which are indicative of the semantic relationship between the sentence they occur in and the respective previous sentence. We then group these linking phrases into three classes based on the type of relationship indicated by each of them. The linking phrases and their assigned class can be seen in Table 2. We select the sentences which start with any of these phrases from each paper and include them in our dataset as hypotheses or second sentences; we include their respective preceding sentences as the premises or first sentences. Each sentence pair is labeled based on the class assigned to the linking phrase present in the second sentence, e.g., if the second sentence starts with "In contrast", the sentence pair is labeled as CONTRASTING. After assigning the labels, we delete the linking phrases from the second sentence of each pair to ensure that the models cannot get any clues of the ground truth labels just by looking at them. We also pair a large number of randomly selected sentences for our NEUTRAL class using three approaches:

- BOTHRAND: Two completely random sentences which do not contain any linking phrases are extracted (both from the same paper) and are paired together.
- FIRSTRAND: First sentence is random; second sentence is selected randomly from the other three classes (both from the same paper).
- SECONDRAND: Second sentence is random; first sentence is selected randomly from the other three classes (both from the same paper).

Our choice for including the last two approaches above was to make the dataset more challenging.

3.3 Benchmark Evaluation Sets Creation

To create our development and test sets, we start by extracting and labeling sentence pairs using the same distant supervision approach described in the previous section from the papers published in 2020 which are available in the ACL anthology. We then manually annotate a subset of these sentence pairs in order to make SCINLI a suitable benchmark for evaluation. The annotation process is completed in two steps, as described below.

First, we manually clean the data by filtering out the examples which contain too many mathematical terms and by completing the sentences that are broken due to erroneous PDF extraction by looking at the papers they are from. The second step of the annotation process is conducted in an

[2] github.com/kermitt2/grobid

Dataset	#Examples			#Words		'S' parser			
	Train	Dev	Test	Prem.	Hyp.	Prem.	Hyp.	Overlap	Agrmt.
SNLI	550,152	10,000	10,000	14.1	8.3	74.0%	88.9%	52.97%	89.0%
MNLI	392,702	20,000	20,000	22.3	-	91.0%	98.0%	-	88.7%
SICK	4,500	-	4,927	9.76	9.57	-	-	64.85%	84.0%
SciTail	23,596	1,304	2,126	10.79	10.28	89.5%	99.1%	54.84%	-%
SciNLI									
+Contrasting	25,353	500	1,000	27.41	24.50	97.3%	97.4%	31.33%	91.6%
+Reasoning	25,353	500	1,000	28.25	24.32	97.5%	97.7%	32.75%	74.6%
+Entailment	25,353	500	1,000	27.08	28.90	96.9%	95.9%	32.98%	82.3%
+Neutral	25,353	500	1,000	26.76	26.02	95.3%	95.6%	23.18%	94.7%
SciNLI Overall	101,412	2,000	4,000	27.38	25.93	96.8%	96.7%	30.06%	85.8%

Table 3: Comparison of key statistics of SciNLI with other related datasets.

iterative fashion. In each iteration, we randomly sample a balanced subset from the cleaned set of examples created in the previous step and present the sentence pair from each example to three expert annotators. To avoid a performance ceiling due to lack of context, the annotators are instructed to label each example based only on the two sentences in each example. If the label is not clear from the context available in the two sentences, the instruction is to label them as unclear. The label with the majority of the votes from annotators is then chosen as the gold label. No gold label is assigned to the examples ($\approx 5\%$) which do not have a majority vote. The examples for which the gold label agrees with the label assigned based on the linking phrase are selected to be in our **benchmark evaluation set**. We continue the iterations of sampling a balanced set of examples and annotating them until we have at least $1,500$ examples from each class in the benchmark evaluation set. In total, $8,044$ sentence pairs—$2,011$ from each class are annotated among which $6,904$ have an agreement between the gold label and the label assigned based on the linking phrase. Therefore, these 6904 examples are selected to be in the benchmark evaluation set. The percentage of overall agreement and the class-wise agreement between the gold labels and the labels assigned based on the linking phrases are reported in the last column of Table 3. The Fleiss-k score among the annotators is 0.62 which indicates that the agreement among the annotators is substantial (Landis and Koch, 1977).

We randomly select 36% of the papers in our benchmark evaluation set to be in our development set and the rest of the papers are assigned to the test set. This is done based on our decision to have at least 500 samples from each class in the development set and 1000 samples from each class in the test set. Splitting the dataset into train, development and test sets *at paper level instead of sentence pair level* is done to prevent any information leakage among the data splits caused by sentences from one paper being in more than one split.

3.4 Data Balancing

Because of the differences in the frequency of occurrence of the linking phrases related to different classes, our initial dataset was unbalanced in all three splits. In contrast, the examples in the related datasets such as SNLI (Bowman et al., 2015) and MNLI (Williams, Nangia, and Bowman, 2018) are almost equally distributed across their classes. Therefore, for a fair comparison, we balance our dataset by downsampling the top three most frequent classes to the size of the least frequent class in each split. We can see the number of examples in each class of our SciNLI dataset in Table 3.

3.5 Data Statistics

A comparison of key statistics of SciNLI with four related datasets is also shown in Table 3.

Dataset Size Although the total size of our dataset is smaller than SNLI and MNLI, SciNLI is still large enough to train and evaluate deep learning based NLI models.

Sentence Lengths From Table 3, we can see that the average number of words in both premise and hypothesis is higher in SciNLI compared with the other datasets. This reflects the fact that sentences used in scientific articles tend to be longer than the sentences used in everyday language.

Sentence Parses Similar to the related datasets, we parse the sentences in SciNLI by using the Stanford PCFG Parser (3.5.2) (Klein and Manning,

Dataset	F1	Acc
SICK	63.54	64.86
SNLI	80.61	80.74
MNLI Dev		
-Matched	65.39	65.70
-Mismatched	64.75	65.01
SciTail	71.18	72.29
SciNLI	60.98	61.38

Table 4: The Macro F1 (%) and Accuracy (%) of the BiLSTM model on different datasets.

2003). We can see that $\approx 97\%$ of both first and second sentences have parses with an 'S' root which is higher than the sentences in SNLI and very competitive with the other datasets. This illustrates that most of our sentences are syntactically complete.

Token Overlap We report the average percentage of tokens occurring in hypotheses which overlap with the tokens in their premises (Table 3). We observe that the overlap percentage in SciNLI is much lower compared to the other datasets. Therefore, our dataset has low surface-level lexical patterns revealing the relationship between sentences.

4 SciNLI Evaluation

We evaluate our dataset by performing three sets of experiments. First, we aim to understand the difficulty level of SciNLI compared to related datasets (§4.1). Second, we investigate a lexicalized classifier to test whether simple similarity based features can capture the particularities of our relations and potentially perform well on our dataset (§4.2). Third, we experiment with traditional machine learning models, neural network models and transformer based pre-trained language models to establish strong baselines (§4.3).

4.1 SciNLI vs. Related Datasets

To evaluate the difficulty of SciNLI, we compare the performance of a BiLSTM (Hochreiter and Schmidhuber, 1997) based classifier on our dataset and four related datasets: SICK, SNLI, MNLI and SciTail. The architecture for this model is similar to the BiLSTM model used by Williams, Nangia, and Bowman (2018). Precisely, the sentence level representations S_1 and S_2 are derived by sending the embedding vectors of the words in each of the sentences in a pair through two separate BiLSTM layers and averaging their hidden states. The context vector S_c is calculated using the following equation:

$$S_c = [S_1, S_2, S_1 \odot S_2, S_1 - S_2] \quad (1)$$

Here, the square brackets denote a concatenation operation of vectors and $\odot$ and $-$ are element-wise multiplication and subtraction operators, respectively. S_c is sent through a linear layer with Relu activation which is followed by a softmax layer to obtain the final output class.

Implementation details We pre-process the input sentences by tokenizing and stemming them using the NLTK tokenizer[3] and Porter stemmer,[4] respectively. Any stemmed token which occurs less than two times in the training set is replaced with an [UNK] token. We use 300D Glove embeddings (Pennington, Socher, and Manning, 2014) to represent the tokens which are allowed to be updated during training. The hidden size for the BiLSTM models is 300. The batch size is set at 64 and the models are trained for 30 epochs where we optimize a cross-entropy loss using Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 0.001. We employ early stopping with a patience size 10 where the Macro F1 score of the development set is used as the stopping criteria. Since SICK does not have a development split, we randomly select 10% of its training examples to be used as the development set. Similarly, since MNLI does not have a publicly available test split, we consider its development split as the test split and we randomly select $\approx 10,000$ samples from the training set to be used as the development set.

We can see the performance of this model on different datasets in Table 4. We find the following:

SciNLI is more challenging than other related datasets. The BiLSTM model shows a much lower performance for SciNLI compared with the other datasets. These results indicate that the task our dataset presents is more challenging compared to other datasets. As we have seen in Table 3, there is a substantial amount of discrepancy in sentence lengths between SciNLI and the other datasets. The longer sentences in our dataset make it harder for the models to retain long distance dependencies, which result in lower performance. Furthermore, our dataset has low surface-level lexical cues and exhibits complex linguistic patterns that require a model to be less reliant on lexical cues but instead learn deep hidden semantics from text.

[3] https://www.nltk.org/api/nltk.tokenize.html

[4] https://www.nltk.org/howto/stem.html

Features	SICK		SciNLI	
	F1	Acc	F1	Acc
UNIGRAMS	33.32	51.39	40.96	41.28
BIGRAMS	33.02	50.90	32.04	32.57
UNIGRAM & BIGRAM	34.52	49.69	39.35	39.52
FEATURES 1-3	66.68	71.86	35.75	38.15
ALL FEATURES	66.22	72.03	47.01	47.78

Table 5: The Macro F1 (%) and Accuracies (%) of the lexicalized classifier on SICK and SCINLI.

4.2 Lexical Similarity vs. Semantic Relationship

To verify that the examples in our dataset cannot be classified based only on syntactic and lexical similarities, we explore a simple lexicalized classifier similar to (Bowman et al., 2015). We train a classifier using different combinations of the following features: (1) the second sentence's BLEU (Papineni et al., 2002) score with respect to the first sentence with an n-gram range of 1 to 4; (2) the difference in length between the two sentences in a pair; (3) overlap of all words, just nouns, verbs, adjectives, or adverbs - both the actual number and the percentage over possible overlaps; and (4) unigrams and bigrams from the second sentence as indicator features. We compare the performance of these models on our dataset and the SICK dataset because given the small size of SICK, this is especially suitable for this kind of models. The results can be seen in Table 5. We observe the following:

Semantic understanding is required to perform well on SCINLI. The lexicalized model fails to achieve satisfactory results on SCINLI even when all features are combined. Both Macro F1 and accuracy are much lower for our dataset than SICK. This means that without actually understanding the content in the sentences in SCINLI, a model cannot successfully predict their relationship.

4.3 SCINLI Baselines

To establish baselines on our dataset, we consider three types of models: a traditional machine learning model, neural network models, and pre-trained language models.

Traditional Machine Learning Model We consider the lexicalized classifier using all four features described in §4.2 as a baseline on our dataset.

Neural Network Models We experiment with three neural models to get the sentence level representations for each sentence in a pair: (a) **BiLSTM** - word embeddings are sent through a BiLSTM layer and the hidden states are averaged; (b) **CBOW** - word embedding vectors are summed; (c) **CNN** - 64 convolution filters of widths [3, 5, 9] on the word embeddings are applied, the outputs of which are mean pooled to get a single vector representation from the filters of each of the three widths. These three vectors are then concatenated to get the sentence level representation.

For all three models, the sentence level representations are combined as in Eq. 1. The obtained representations are first sent through a linear layer with Relu activation followed by softmax for classification (i.e., project them with a weight matrix $\mathbf{W} \in \mathbb{R}^{d \times 4}$). The hyperparameters and other implementation details are the same as for the BiLSTM model described in §4.1.

Pre-trained Language Models We fine-tune four transformer based pre-trained language models: (a) **BERT** (Devlin et al., 2019) - pre-trained by masked language modeling (MLM) on BookCorpus (Zhu et al., 2015) and Wikipedia; (b) **SciBERT** (Beltagy, Lo, and Cohan, 2019) - a variant of BERT pre-trained with a similar procedure but exclusively on scientific text; (c) **RoBERTa** (Liu et al., 2019b) - an extension of BERT which was pre-trained using dynamic masked language modeling, i.e., unlike BERT, different words were masked in each epoch during training. It was also trained for a longer period of time on a larger amount of text compared with BERT; and (d) **XLNet** (Yang et al., 2019) - pre-trained with a "Permutation Language Modeling" objective instead of MLM. We employ the base variants of each of these models using the huggingface transformers library. The input sequence for these models is derived by concatenating the two sentences in a pair with a `[SEP]` token in between. The `[CLS]` token is then projected with a weight matrix $\mathbf{W} \in \mathbb{R}^{d \times 4}$ by sending it as the input to a softmax layer to get the output class. We fine-tune each transformer based model for 5 epochs where we minimize the cross-entropy loss using Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of $2e-5$. Early stopping with a patience size 2 is employed.

The experiments are run on a single Tesla V10 GPU. The transformer based models took approximately four hours to train and the traditional machine learning and neural network models were trained in less than one hour. We run each experiment three times with different random seeds and

	CONTRASTING	REASONING	ENTAILMENT	NEUTRAL	Macro F1	Acc
Lexicalized	50.28 ± 0.00	37.18 ± 0.00	44.82 ± 0.00	55.77 ± 0.00	47.01 ± 0.00	47.78 ± 0.00
CBOW	54.62 ± 2.17	50.54 ± 1.75	52.33 ± 3.42	49.25 ± 0.18	51.68 ± 0.48	51.78 ± 0.53
CNN	63.73 ± 1.59	58.86 ± 1.17	62.66 ± 0.76	56.40 ± 0.97	60.41 ± 0.86	60.53 ± 0.85
BiLSTM	63.93 ± 0.53	57.32 ± 2.05	64.01 ± 0.56	59.25 ± 0.60	61.12 ± 0.15	61.32 ± 0.08
BERT	77.46 ± 0.30	71.74 ± 0.82	75.09 ± 0.13	76.47 ± 1.70	75.19 ± 0.35	75.17 ± 0.39
SciBERT	80.30 ± 0.60	74.18 ± 0.33	75.90 ± 1.47	79.76 ± 0.25	$77.53^{*} \pm 0.49$	$77.52^{*} \pm 0.49$
RoBERTa	81.18 ± 0.77	74.22 ± 0.81	77.99 ± 0.52	78.86 ± 0.61	$78.06^{*} \pm 0.39$	$78.12^{*} \pm 0.33$
XLNet	81.53 ± 0.30	75.95 ± 0.94	77.63 ± 0.38	77.63 ± 0.68	$\mathbf{78.18^{*} \pm 0.06}$	$\mathbf{78.23^{*} \pm 0.12}$

Table 6: The Macro F1 scores (%) and accuracies (%) of our baseline models on SCINLI along with individual F1 scores on four classes. Here, an asterisk indicates that there is a statistically significant difference between the models in the third block of the table and BERT according to a paired T-test with $\alpha = 0.05$. The three models in the third block shows statistically indistinguishable results. The best Macro F1 and accuracy are in bold.

report the average and standard deviation of the F1 scores for each of the four classes, their Macro average and overall accuracy in Table 6. Our findings are discussed below.

Transformer based models consistently outperform the traditional models The transformer based models have a very high performance gap with the traditional lexicalized and neural models. Their better performance can be attributed to their superior design for capturing the language semantics and their pre-training on large amounts of texts.

More sophisticated pre-training methods lead to better performance RoBERTa and XLNet are created by addressing different limitations of BERT. Both of these models show a better performance than BERT on our dataset. Therefore, the progress made in these two models for better NLU capability is reflected by the results on SCINLI. This proves that SCINLI can be used as an additional resource for tracking the progress of NLU.

Pre-training on domain specific text helps to improve classification performance The results show that SciBERT consistently outperforms BERT on SCINLI. This is because unlike BERT, SciBERT was pre-trained exclusively on scientific text. Hence, it has a better capability to understand the text in the scientific domain. We see that RoBERTa and XLNet show slightly better performances than SciBERT despite being pre-trained on non-scientific text, just like BERT. However, it should be noted that these differences in performance are not statistically significant. Moreover, both RoBERTa and XLNet were created by modifying the training procedure of BERT to further improve the performance, whereas SciBERT is just a plain BERT model pre-trained on scientific text. Even without any modifications to the training procedure, SciBERT is able to perform similarly to these models proving the advantage of pre-training on domain specific text and suitability of our dataset for evaluating scientific NLI models.

Model		SCINLI F1	SCINLI Acc
BERT	BOTH SENTENCES	75.36	75.37
	ONLY 2^{nd} SENTENCE	54.56	55.40
SciBERT	BOTH SENTENCES	77.66	77.60
	ONLY 2^{nd} SENTENCE	58.16	58.80

Table 7: Performance comparison on SCINLI when both sentences are concatenated vs. when only second sentence is used as the input.

5 Analysis

Research has shown that some stylistic and annotation artifacts are present (only in the hypotheses) in NLI datasets created using crowdsource annotators (Gururangan et al., 2018). To verify that the models do not learn similar spurious patterns in our dataset and predict the labels without understanding the semantic relation between the sentences, we start our analysis by experimenting with only the second sentence as the input to BERT and SciBERT models. Next, to intuitively understand the errors made by the models, we perform a qualitative analysis of the predictions made by the SciBERT model on 100 randomly selected examples from our test set. Finally, we show that the NEUTRAL examples extracted with FIRSTRAND and SECONDRAND approaches are harder to classify than the examples extracted with BOTHRAND.

Spuriosity Analysis A comparison between the *only second sentence* models and the models with both sentences concatenated as the input can be seen in Table 7. Clearly, as we can see from the

First Sentence	Second Sentence	True Label	Predicted Label
Multiple studies of BERT concluded that it is considerably overparametrized.	it is possible to ablate elements of its architecture without loss in performance or even with slight gains (Kovaleva et al., 2019;Michel et al., 2019;Voita et al., 2019).	ENTAILMENT	CONTRASTING
Upon further investigation, we find that experiments which use probabilities with image based features have an inter-quartile range of 0.05 and 0.1 for EBG and BLOG respectively whereas for experiments using probabilities with binning based features, this range is 0.32 for both datasets.	inter-quartile range for experiments using ranks with image based features is 0.08 and 0.05 for EBG and BLOG whereas for experiments using ranks with binning based features, this range is 0.49 and 0.42 respectively.	CONTRASTING	NEUTRAL

Table 8: Examples of errors made by SciBERT on SCINLI.

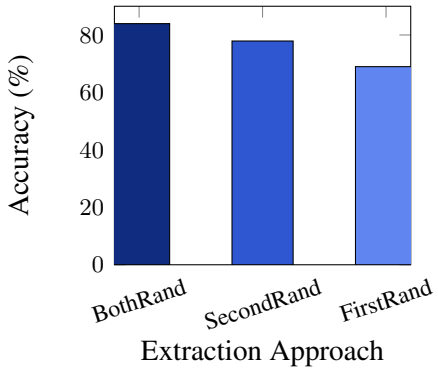

Figure 1: Extraction approach vs. accuracy of SciBERT on the NEUTRAL pairs of SCINLI test set.

table, there is a substantial amount of performance decrease when only the second sentence is used as input. Therefore, in order to perform at the optimal level, both sentences are required for the models to make the correct inference by learning the semantic relation between them.

Qualitative Error Analysis We find that a major reason behind the wrong predictions is a lack of domain specific knowledge. For example, in the first sentence pair in Table 8, without the domain knowledge that the number of parameters in a model affects the performance, one will not be able to make the correct inference. We also find that the model is prone to making mistakes for longer sentences. This issue is exemplified by the second sentence pair in Table 8.

Neutral Class Performance Analysis We can see a plot of the accuracy shown by SciBERT on NEUTRAL pairs of our test set extracted with different approaches in Figure 1. Indeed, the examples in which one sentence comes from one of the other three classes are harder to classify.

6 Conclusion & Future Directions

In this paper, we introduced SCINLI, the first natural language inference dataset on scientific text created with our novel data annotation method. We manually annotated a large number of examples to create our benchmark test and development sets. Our experiments suggest that SCINLI is harder to classify than existing NLI datasets and deep semantic understanding is necessary for a model to perform well. We establish strong baselines and show that our dataset can be used as a challenging benchmark to evaluate the progress of NLU models. In the future, we will leverage knowledge bases to improve the models' ability to understand scientific text. We make our code and the SCINLI dataset available to further research in scientific NLI.

Acknowledgements

This research is supported by NSF CAREER award 1802358 and NSF CRI award 1823292 to Cornelia Caragea. Any opinions, findings, and conclusions expressed here are those of the authors and do not necessarily reflect the views of NSF. We thank AWS for computing resources. We also thank our anonymous reviewers for their constructive feedback, which helped improve our paper.

References

Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3615–3620. Hong Kong, China: Association for Computational Linguistics.

Bird, S.; Dale, R.; Dorr, B. J.; Gibson, B.; Joseph,

M. T.; Kan, M.-Y.; Lee, D.; Powley, B.; Radev, D. R.; and Tan, Y. F. 2008. The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *Proc. of the 6th International Conference on Language Resources and Evaluation Conference (LREC'08)*, 1755–1759.

Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 632–642. Lisbon, Portugal: Association for Computational Linguistics.

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*.

Dagan, I.; Glickman, O.; and Magnini, B. 2005. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, 177–190. Springer.

Dagan, I.; Glickman, O.; and Magnini, B. 2006. The PASCAL Recognising Textual Entailment Challenge. In Quiñonero-Candela, J.; Dagan, I.; Magnini, B.; and d'Alché Buc, F., eds., *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, 177–190. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-540-33428-6.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Do, Q.; Chan, Y. S.; and Roth, D. 2011. Minimally Supervised Event Causality Identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 294–303. Edinburgh, Scotland, UK.: Association for Computational Linguistics.

Dunietz, J.; Levin, L.; and Carbonell, J. 2017. The BECauSE Corpus 2.0: Annotating Causality and Overlapping Relations. In *Proceedings of the 11th Linguistic Annotation Workshop*, 95–104. Valencia, Spain: Association for Computational Linguistics.

Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S.; and Smith, N. A. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 107–112. New Orleans, Louisiana: Association for Computational Linguistics.

Hall, D.; Jurafsky, D.; and Manning, C. D. 2008. Studying the history of ideas using topic models. In *Proceedings of the conference on empirical methods in natural language processing*, 363–371. Association for Computational Linguistics.

Hobbs, J. R. 1978. Why is discourse coherent. Technical report, SRI INTERNATIONAL MENLO PARK CA.

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

Jernite, Y.; Bowman, S. R.; and Sontag, D. 2017. Discourse-based objectives for fast unsupervised sentence representation learning. *arXiv preprint arXiv:1705.00557*.

Khot, T.; Sabharwal, A.; and Clark, P. 2018. SciTaiL: A Textual Entailment Dataset from Science Question Answering. In *AAAI*, volume 17, 41–42.

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

Klein, D.; and Manning, C. D. 2003. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 423–430. Sapporo, Japan: Association for Computational Linguistics.

Kuhn, T. S. 2012. *The structure of scientific revolutions*. University of Chicago press.

Landis, J. R.; and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *biometrics*, 159–174.

Li, Z.; Ding, X.; Liu, T.; Hu, J. E.; and Van Durme, B. 2020. Guided Generation of Cause and Effect. In Bessiere, C., ed., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, 3629–3636. International Joint Conferences on Artificial Intelligence Organization. Main track.

Liu, X.; He, P.; Chen, W.; and Gao, J. 2019a. Multi-Task Deep Neural Networks for Natural Language Understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4487–4496. Florence, Italy: Association for Computational Linguistics.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Luukkonen, T. 1992. Is scientists' publishing behaviour rewardseeking? *Scientometrics*, 24: 297–319.

Marelli, M.; Menini, S.; Baroni, M.; Bentivogli, L.; Bernardi, R.; and Zamparelli, R. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 216–223. Reykjavik, Iceland: European Language Resources Association (ELRA).

Mintz, M.; Bills, S.; Snow, R.; and Jurafsky, D. 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*, 1003–1011. Suntec, Singapore: Association for Computational Linguistics.

Nie, A.; Bennett, E.; and Goodman, N. 2019. DisSent: Learning Sentence Representations from Explicit Discourse Relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4497–4510. Florence, Italy: Association for Computational Linguistics.

Nie, Y.; Williams, A.; Dinan, E.; Bansal, M.; Weston, J.; and Kiela, D. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4885–4901. Online: Association for Computational Linguistics.

Oh, J.-H.; Torisawa, K.; Hashimoto, C.; Sano, M.; De Saeger, S.; and Ohtake, K. 2013. Why-Question Answering using Intra- and Inter-Sentential Causal Relations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1733–1743. Sofia, Bulgaria: Association for Computational Linguistics.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Prasad, R.; Dinesh, N.; Lee, A.; Miltsakaki, E.; Robaldo, L.; Joshi, A.; and Webber, B. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA).

Pruksachatkun, Y.; Phang, J.; Liu, H.; Htut, P. M.; Zhang, X.; Pang, R. Y.; Vania, C.; Kann, K.; and Bowman, S. R. 2020. Intermediate-Task Transfer Learning with Pretrained Language Models: When and Why Does It Work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5231–5247. Online: Association for Computational Linguistics.

Radev, D. R.; Muthukrishnan, P.; and Qazvinian, V. 2009. The ACL Anthology Network. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries (NLPIR4DL)*, 54–61. Suntec City, Singapore: Association for Computational Linguistics.

Radinsky, K.; Davidovich, S.; and Markovitch, S. 2012. Learning causality for news events prediction. In *Proceedings of the 21st international conference on World Wide Web*, 909–918.

Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, 3266–3280.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 353–355. Brussels, Belgium: Association for Computational Linguistics.

Webber, B.; Knott, A.; Stone, M.; and Joshi, A. 1999. Discourse relations: A structural and presuppositional account using lexicalised TAG. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 41–48.

Williams, A.; Nangia, N.; and Bowman, S. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1112–1122. Association for Computational Linguistics.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, 5753–5763.

Zellers, R.; Bisk, Y.; Schwartz, R.; and Choi, Y. 2018. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 93–104. Brussels, Belgium: Association for Computational Linguistics.

Zhu, Y.; Kiros, R.; Zemel, R.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, 19–27.