Natural Language Inference in Context — Investigating Contextual Reasoning over Long Texts

Hanmeng Liu,[1] **Leyang Cui,**[1] **Jian Liu,**[2] **Yue Zhang**[3]
[1] Zhejiang University, [2] Fudan University, [3] Westlake University
liuhanmeng@zju.edu.cn, cuileyang@westlake.edu.cn, jianliu17@fudan.edu.cn, yue.zhang@wias.org.cn

Abstract

Natural language inference (NLI) is a fundamental NLP task, investigating the entailment relationship between two texts. Popular NLI datasets present the task at sentence-level. While adequate for testing semantic representations, they fall short for testing contextual reasoning over long texts, which is a natural part of the human inference process. We introduce ConTRoL, a new dataset for **ConT**extual **R**easoning **o**ver **L**ong Texts. Consisting of 8,325 expert-designed "context-hypothesis" pairs with gold labels, ConTRoL is a passage-level NLI dataset with a focus on complex contextual reasoning types such as logical reasoning. It is derived from competitive selection and recruitment test (verbal reasoning test) for police recruitment, with expert level quality. Compared with previous NLI benchmarks, the materials in ConTRoL are much more challenging, involving a range of reasoning types. Empirical results show that state-of-the-art language models perform by far worse than educated humans. Our dataset can also serve as a testing-set for downstream tasks like checking the factual correctness of summaries.

Introduction

Natural languages are powerful tools for reasoning. In NLP, natural language inference (NLI) has attracted surging research interests (Bowman et al. 2015; Williams, Nangia, and Bowman 2018; Bhagavatula et al. 2020). The task is to determine whether a hypothesis h can reasonably be inferred from a premise p. Thanks to the generalizability of the NLI framework (i.e., nearly all questions about meaningfulness in language can be reduced to questions of entailment and contradiction in context), NLI can serve as a proxy to general tasks such as natural language understanding (NLU). As a result, the NLI task is constantly employed as a testing ground for learning sentence representation as well as evaluating language models, with the expectation of benefiting downstream applications.

Large-scale NLI datasets have been collected via crowdsourcing. Existing benchmarks (Bowman et al. 2015; Williams, Nangia, and Bowman 2018; Dagan, Glickman, and Magnini 2005; Khot, Sabharwal, and Clark 2018a) handle the task at the sentence-level, generating labelled sentence pairs by probing into the essence of lexical and compositional semantics. These benchmarks explore rich features of sentence meaning, testing various aspects of semantic representation. With the advance of contextualized embeddings such as BERT (Devlin et al. 2019), pre-trained language models achieve competitive results. The state-of-the-art models can even reach human-level performance.

P: Ten new television shows appeared during the month of September. Five of the shows were sitcoms, three were hour-long dramas, and two were news-magazine shows. By January, only seven of these new shows were still on the air. Five of the shows that remained were sitcoms.

H1: *At least one of the shows that were cancelled was an hour-long drama.*
Entailment ✓ Contradiction Neutral

H2: *There is no hour-long drama remained on the air.*
Entailment **Contradiction** ✓ Neutral

H3: *Television viewers prefer sitcoms over hour-long dramas.*
Entailment Contradiction **Neutral** ✓

Figure 1: An example of the ConTRoL dataset. (✓ indicates the correct answer.)

Contextual reasoning is essential to the process of human cognition, where inference is made based on contextual information and a collection of facts (Giunchiglia 1992). Inferring hidden facts from context is an indispensable element of human language understanding. Contextual reasoning is typically performed on the passage level, where multiple steps may be necessary for inferring facts from given evidences. It has been investigated by NLP tasks such as machine reading (Lai et al. 2017; Sun et al. 2019a), retrieval-based dialogue (Wu et al. 2017). However, dominant NLI benchmarks (Bowman et al. 2015; Williams, Nangia, and Bowman 2018) investigate the relationship of two sentences, with relatively less attention being payed to the exploration of grounded logical inference (Bhagavatula et al. 2020; Clark, Tafjord, and Richardson 2020).

We investigate contextual reasoning for NLI by making a dataset that consists of 8,325 instances. One example is shown in Figure 1. In this example, the premise consists of several facts concerning a set of shows, which can serve as a context for evidence integration and reasoning. The truthfulness of the hypotheses are determined by reasoning over multiple sentences. Various types of contextual reason-

Dataset	Task	Reasoning	Context	Source
SQuAD (Rajpurkar et al. 2016)	Reading Comprehension	✓	Passage	Wikipedia
WIKIHOP (Welbl, Stenetorp, and Riedel 2018)	Reading Comprehension	✓	Document	Wikipedia
HOTPOTQA (Yang et al. 2018)	Reading Comprehension	✓	Document	Wikipedia
Cosmos QA (Huang et al. 2019)	Reading Comprehension	✓	Passage	Webblog
Social IQA (Sap et al. 2019)	Reading Comprehension	✗	Sentence	Social
WINOGRANDE (Sakaguchi et al. 2020)	Coreference Resolution	✗	Sentence	Diverse
CommonsenseQA (Talmor et al. 2019)	Reading Comprehension	✗	Sentence	Diverse
MuTual (Cui et al. 2020)	Next Utterance Prediction	✓	Dialogue	Exam
ReClor (Yu et al. 2020)	Reading Comprehension	✓	Passage	Exam
LogiQA (Liu et al. 2020a)	Reading Comprehension	✓	Passage	Exam
RTE (Dagan, Glickman, and Magnini 2005)	Natural Language Inference	✗	Sentence	Diverse
SNLI (Bowman et al. 2015)	Natural Language Inference	✗	Sentence	Captioning
WNLI (Wang et al. 2018)	Natural Language Inference	✗	Sentence	Fiction
QNLI (Wang et al. 2018)	Natural Language Inference	✗	Sentence	Wikipedia
MultiNLI (Williams, Nangia, and Bowman 2018)	Natural Language Inference	✗	Sentence	Diverse
Dialogue NLI (Welleck et al. 2019)	Natural Language Inference	✗	Sentence	Persona
SciTaiL (Khot, Sabharwal, and Clark 2018a)	Natural Language Inference	✗	Sentence	Science
Adversarial NLI (Nie et al. 2020)	Natural Language Inference	✗	Paragraph	Diverse
AlphaNLI (Bhagavatula et al. 2020)	Natural Language Inference	✗	Sentence	Diverse
ConTRoL	Natural Language Inference	✓	Passage	Exam

Table 1: Comparison between our dataset and existing benchmarks. "Reasoning" refers to contextual reasoning.

ing are considered in the dataset, with more examples being shown in Figure 2 We name our open-domain dataset **ConT**extual **R**easoning **o**ver **L**ong Texts (ConTRoL), which is a passage-level natural language inference dataset with gold label data. It differs from the existing NLI datasets in the following three main aspects: (1) the materials are sourced from verbal reasoning exams which are expert-designed rather than crowdsouced; (2) they inspect the abilities of various reasoning types; (3) the contexts are more complex than previous datasets with longer spans.

We evaluate the state-of-the-art NLI models to establish baseline performances for ConTRoL. Experimental results demonstrate a significant gap between machine and human ceiling performance. Detailed analysis is given to shed light on future research. Our dataset and results are released at https://github.com/csitfun/ConTRoL-dataset.

Related Work

Natural Language Inference

The task of text entailment was introduced in the PASCAL Recognizing Textual Entailment (RTE) challenges (Dagan, Glickman, and Magnini 2005), which deals with relationship of sentence pairs. On the third RTE challenge (Giampiccolo et al. 2007), a very limited number of longer texts with multiple sentences were incorporated for more comprehensive scenarios. This shares a similar idea to our work, yet the challenge does not give multi-sentence materials at scale for detailed study.

Recently, the most widely used NLI benchmarks include the Stanford Natural Language Inference (SNLI) dataset (Bowman et al. 2015), and the subsequently expanded MultiNLI (Williams, Nangia, and Bowman 2018), bringing sentences of various genres into the original SNLI. MultiNLI is included in the GLUE benchmark (Wang et al. 2018) and is widely used in evaluating language models' performance. Other NLI datasets include the Question-answering NLI (QNLI) (Wang et al. 2018), the Winograd NLI (WNLI) (Wang et al. 2018), the SciTail (Khot, Sabharwal, and Clark 2018b) etc., which focuses on different aspects of knowledge. While all the above datasets are on the sentence-level, we investigate NLI for long texts.

Dialogue NLI (Welleck et al. 2019) features a persona-based dialogue structure for making inference on the current utterance based on previous dialogue history. Similar to our dataset, discourses involve multi-sentence context as premises. However, they do not consider relationships that require more than two sentences to express, nor is logical reasoning explored.

The multiple premise entailment (MPE) task (Lai, Bisk, and Hockenmaier 2017) is a variant of the entailment task where each hypothesis is paired with a set of independently written premise sentences. It is similar to our dataset in the spirit of using longer contexts, while the premise sentences in the MPE task are derived from image captioning. The describing sentences are fragmented and not in order. The dataset exams different types of semantic phenomena that are useful for inference, which is different from our dataset fundamentally. Similarly, Adversarial NLI (Nie et al. 2020) holds the simple intuition that longer contexts lead to harder examples, which coincide with our idea to some extent. The Adversarial NLI dataset is similar to ours in that longer contexts are considered in the premises. However, we differ in context length and reasoning types. The context of our dataset is much longer and with *multiple* paragraphs being involved. In contrast, Adversarial NLI has single-paragraph contexts only. In addition, it does not test logical reasoning, which is the main focus of ConTRoL. To our knowledge, we

are the first to introduce a passage-level NLI dataset requiring comprehensive grounded logical reasoning.

AlphaNLI (Bhagavatula et al. 2020) explores the problem of abductive reasoning. It asks for the most plausible explanation given observations from *two* narrative contexts. Similar to our dataset, investigating the nature of human reasoning is the target of AlphaNLI. However, the AlphaNLI challenge resembles the classical formulation of abductive reasoning, which is different from the reasoning types we are focused on. In addition, both premises (i.e., observation contexts) in AlphaNLI are written in single sentences. In contrast, our dataset consists of multi-paragraph premises.

Clark, Tafjord, and Richardson (2020) investigated deductive reasoning by synthesizing a dataset related to NLI. The input is a set of facts and a set of rules that explain the facts, which can be viewed as a premise, together with a fact that can be viewed as a hypothesis, and the output is a binary class *true* or *false*. Compared with their dataset, our work differs in two aspects. First, we consider more reasoning types. Second, ConTRoL is a multi-paragraph NLI dataset with human-written inputs.

Contextual Reasoning

Long texts with multiple paragraphs have been explored in reading comprehension. In particular, there have been challenges that examine evidence integration over multiple text passages (Welbl, Stenetorp, and Riedel 2018; Yang et al. 2018; Rajpurkar, Jia, and Liang 2018), and challenges that focus on commonsense reasoning (Talmor et al. 2019; Cui et al. 2020), including social commonsense (Sap et al. 2019) and external knowledge (Huang et al. 2019; Sakaguchi et al. 2020). Different from these datasets, ConTRoL examines more complex contextual reasoning types such as *logical reasoning*.

There have been reading comprehension datasets that examine logical reasoning. LogiQA (Liu et al. 2020a) is sourced from public service exams. It focuses on linguistic reasoning questions typically featured with a question and four possible answers. ReClor (Yu et al. 2020) is a reading comprehension dataset that is sourced from the GMAT and LSAT test. Similar to our dataset, these datasets examine a range of different logical reasoning types. Different from these benchmarks, ConTRoL takes the form of NLI, which is a more fundamental linguistic task and relevant to different downstream tasks. The correlation and differences between exisiting datasets are shown in Table 1.

Dataset

Crowdsourcing has been a widely-adopted practice for developing large-scale NLI datasets (Bowman et al. 2015; Williams, Nangia, and Bowman 2018; Nie et al. 2020). However, producing a high-quality dataset addressing complex logical reasoning can be difficult for crowdsource workers. Annotation artefacts exist in crowdsourced datasets, for the annotation protocols encourage workers to adopt heuristics to generate hypotheses quickly and efficiently (Gururangan et al. 2018). To avoid such issues, we source our dataset from examinations, and in particular senior aptitude tests (verbal reasoning test), which are designed by experts.

	ConTRoL
Construction Method	Exams
Context Type	Passage
# of passages	1,970
# of premise-hypothesis pairs	8,325
# of multi-paragraph	4,171
Avg. length of multi-paragraph	757
# of single-paragraph	4,154
Avg. length of single-paragraph	148
Vocab size (premise)	54,265
Vocab size (hypothesis)	14,323
Avg. premise length	452
Avg. hypothesis length	12
Lexical overlap (Entailment)	4.87%
Lexical overlap (Neutral)	4.19%
Lexical overlap (Contradiction)	5.49%

Table 2: Data statistics of ConTRoL.

Data Collection and Statistics

We collect our data from publicly available online practice tests, which include verbal logical reasoning tests in the Police Initial Recruitment Test (PIRT), verbal reasoning tests used by the Medical College Admission Test (MCAT) and University Clinical Aptitude Test (UCAT), as well as verbal aptitude tests adopted by corporations' employee recruitment & selection online test. Unlike reading comprehension tests, which can be diverse both in question types and options, questions in the original verbal reasoning tests are similar in struct to NLI tests, where a premise and a hypothesis are given, and the answer is a choice from three options: *true*, *false* and *cannot say*. This corresponds to the three-label setting of the NLI task and we can easily convert the three answer choices into ENTAILMENT, CONTRADICTION and NEUTRAL respectively.

The verbal reasoning tests require exam-takers to comprehend meaning and significance, assess logical strength, make valid inference, and identify a valid summary, interpretation or conclusion. The subjects of the passages are drawn from a range of fields, such as current affairs, business, science, the environment, economics, history, meteorology, health and education. The questions are of high quality, advanced in difficulty level, used in exams such as police initial selection and other highly intellectual practices' candidate recruitment.

The detailed statistics of ConTRoL are shown in Table 2. After removing all duplicated questions, we obtain 8,325 context-hypothesis pairs. We also calculate the lexical overlap between context and hypothesis, finding only 4.87% overlap in the ENTAILMENT relationship, and 5.49% in the CONTRADICTION relationship. This suggests that ConTRoL can be difficult to solve by plain lexical matching.

Data Format

The data format of ConTRoL follows existing NLI benchmarks (Bowman et al. 2015; Williams, Nangia, and Bowman 2018), where each instance contains a premise, a hypothesis, and a label from ENTAILMENT, NEUTRAL and CON-

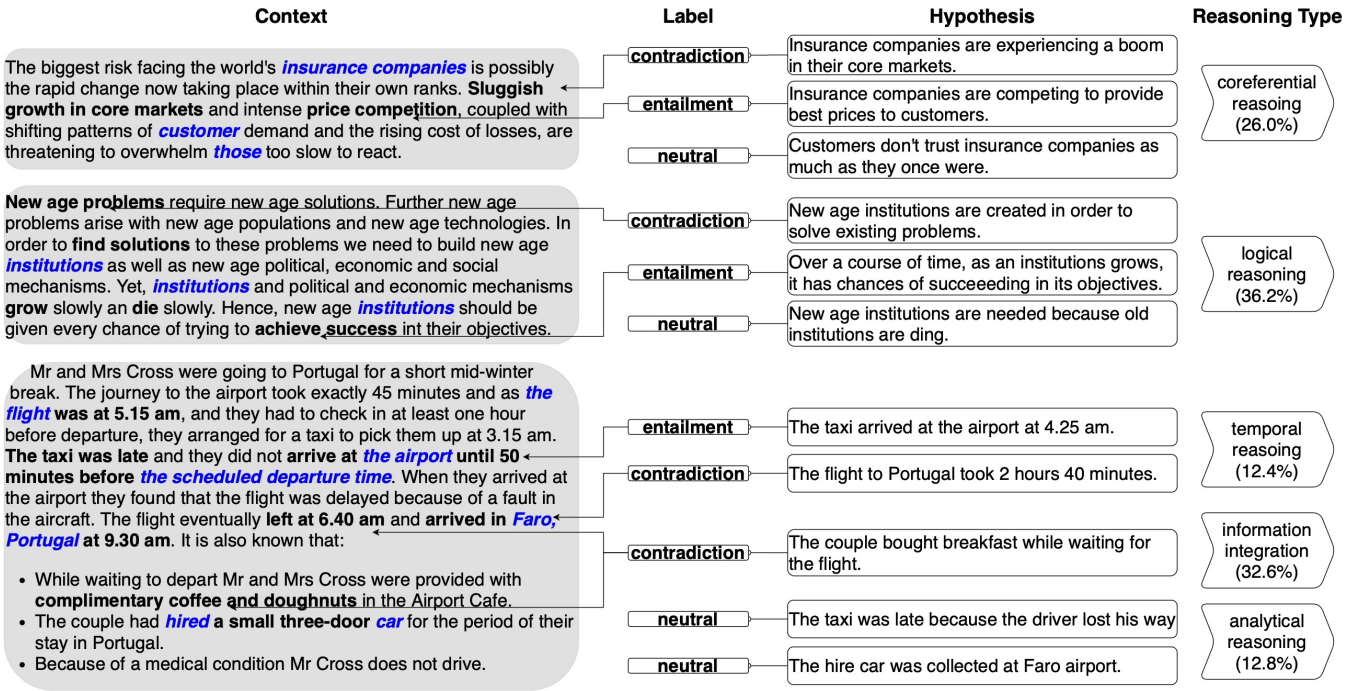

Figure 2: Reasoning types in ConTRoL (Reasoning clues are highlighted in the context).

TRADICTION. Different from existing datasets, the premises are much longer, in one or more paragraphs. In addition, for each premise, three or more hypotheses are given, which is another distinction from former NLI datasets.

Reasoning Types

We manually categorize the test instances by the reasoning type, which can be described as follows:

- **Coreferential Reasoning over Long Texts**
  Coreferential reasoning (Ye et al. 2020) is a form of reasoning over multiple mentions. Long text can accommodate complex relationships between noun phrases, which makes coreferential reasoning crucial for the coherent understanding of texts.
- **Verbal Logical Reasoning**
  Verbal logical reasoning (Liu et al. 2020a) is the ability to examine, analyze, and critically evaluate arguments as they occur in ordinary language. In contrast to formal logical reasoning, most of which uses abstract diagrammatical cues, verbal logical reasoning concerns the logical inference of human language. Deep logical reasoning can be necessary in comprehending long texts.
- **Temporal and Mathematical Reasoning**
  Time and sequential cues of events and requires the ability to reason about time and do the necessary mathematical calculation. Temporal reasoning (Nakhimovsky 1987) is the process of extracting temporal cues and combining them into a coherent temporal view. Various types of temporal information can be found in ConTRoL.
- **Information Integration over Paragraphs**
  Multi-step reasoning (Liu et al. 2020b; Wang et al. 2019; Welbl, Stenetorp, and Riedel 2018) is the ability to retrieve and combine information from multiple paragraphs or multiple documents. For each hypothesis, readers find the most relevant paragraphs in a premise through an iterative (multi-step) process between the contexts and the hypotheses.
- **Analytical Reasoning**
  Analytical reasoning (Williams et al. 2019) is the ability for problem solving to consider a group of facts and rules, and determine the validity of new facts. The fact sets are based on a single or multiple paragraphs, reflecting the kinds of detailed analyses of relationships and sets of constraints. Reasoning is based on what is required given the scenario, what is permissible given the scenario, and what is prohibited given the scenario.

Examples of the above reasoning types can be found in Figure 2. Among all the reasoning types, logical reasoning takes 36.2% of all the test instances, followed by information integration, which takes 32.6%. The proportion of coreferential reasoning, analytical reasoning and temporal reasoning are 26.0%, 12.8% and 12.4%, respectively. It is also worth noticing that one context-hypothesis pair may contain more than one reasoning type, under which circumstance we take the most significant one into the statistics.

Models

We establish several strong baseline methods using the state-of-the-art pre-trained language models.

	Overall		Entailment			Neutral			Contradiction		
	Acc	Micro F1	P	R	F1	P	R	F1	P	R	F1
Human	87.06	93.15	94.83	95.65	95.24	93.33	91.21	92.26	93.02	90.91	91.95
Ceiling	94.40	97.26	99.16	99.16	99.16	97.72	93.75	95.69	96.09	97.79	96.93
BERT-base	47.39	46.22	43.84	54.40	42.45	39.67	51.07	50.21	41.65	52.68	46.00
BERT-large	50.62	49.49	45.15	59.32	45.96	44.21	53.52	53.19	44.68	56.27	49.31
RoBERTa	45.90	45.67	40.99	51.24	45.38	**47.93**	44.34	45.96	44.19	47.54	45.67
Longformer	49.88	46.22	43.24	58.88	45.64	46.28	54.74	46.81	44.71	56.74	46.22
BART	**56.34**	**54.18**	**50.23**	**67.32**	**49.12**	44.21	**62.99**	**59.57**	**47.03**	**65.09**	**53.85**
BART-NLI	45.02	42.33	39.85	53.49	40.87	43.80	46.79	43.83	41.73	49.92	42.30
BART-NLI-FT	60.95	57.41	62.58	61.54	58.67	42.15	78.29	56.17	50.37	68.91	57.39

Table 3: Experiment results on ConTRoL. BART-NLI indicates training on SNLI, MultiNLI and Adversarial NLI and testing on ConTRoL. BART-NLI-FT indicates BART-NLI followed by a fine-tuning step on ConTRoL.

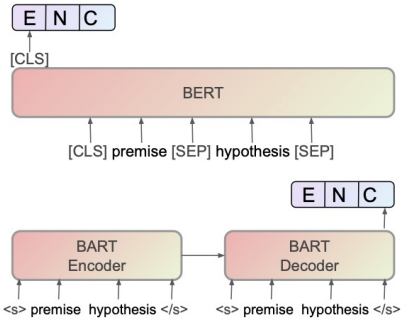

Figure 3: The model structure of BERT and BART. ("E" represents ENTAILMENT, "N" represents NEUTRAL, "C" represents CONTRADICTION

Pre-trained Language Models

BERT (Devlin et al. 2019) is a Transformer-based (Vaswani et al. 2017) language model. During pre-training, BERT uses a masked language modeling objective. The basic idea is to train a model to make use of bidirectional context information for predicting a masked token, so that linguistic knowledge can be collected from large texts. It has been shown that such a language model contains certain degrees of syntactic (Goldberg 2019), semantic (Clark et al. 2019), commonsense (Cui et al. 2020) and logical reasoning (Clark, Tafjord, and Richardson 2020) knowledge.

RoBERTa (Liu et al. 2019) extends BERT using a more dynamic sentence masking method.

Longformer (Beltagy, Peters, and Cohan 2020) Traditional self-attention operation are unable to process long sequences, which scales quadratically with the sequence length. The aforementioned Transformer-based models constrain the input to 512 tokens. To address this limitation, Longformer adopts sliding window attention with global attention to replace the self-attention mechanism in pretrained Transformers.

BART (Lewis et al. 2020) is a denoising autoencoder for pre-training sequence-to-sequence models by combining bidirectional and auto-regressive Transformers.

NLI Model

The NLI model structures of BART and BERT-based are illustrated in Figure 3. For BERT-based models (i.e., BERT, RoBERTa, XLNet and Longformer), following Devlin et al. (2019), given a premise P and a hypothesis h, we concatenate premise-hypothesis pair as a new sequence $[CLS]+p+[SEP]+h+[SEP]$, where [CLS] and [SEP] are special symbol for classification token and separator token. After pre-training model encoding, the last layer's hidden representation from the [CLS] token is fed in an MLP+softmax for classification. For BART, we feed the same sequence to both the encoder and the decoder, using the last hidden state for classification. The class that corresponds to the highest probability is chosen as model prediction.

Implementation Details

We randomly split the dataset into training, development, and test set with the ratio of 8:1:1. All models are trained for 10 epochs. We find hyper-parameters using grid search: batch size $\in \{8, 16, 32\}$ learning rate $\in \{1e-5, 2e-5, 3e-5, 4e-5, 5e-5\}$ and gradient accumulate step $\in \{1, 2, 4\}$. We set the max length to 512 tokens for all models except Longformer, of which 3,000 tokens are the max length we take. Models with the best performance on the development set are used for testing.

Evaluation

Following the NLI benchmark setting (Bowman et al. 2015; Williams, Nangia, and Bowman 2018; Welleck et al. 2019), we employ the overall accuracies as the main evaluation method. Furthermore, to give more detailed analysis, we also calculate precision (P), recall (R) and F1-score (F1) on the ENTAILMENT, NEUTRAL and CONTRADICTION labels.

Human Performance

To measure human performance on the ConTRoL dataset, we randomly select 300 context-hypothesis pairs from the test set. Four testees are recruited. The testees are well educated, two of them are post-graduate students and two of them have PhD degrees. We report the human performance by the mean score and standard deviation. The human ceiling performance is obtained by considering the proportion of questions with at least one correct answer.

Benchmark	# Train	# Test	BERT	SOTA Model	SOTA Performance	Human
MultiNLI	393k	20k	85.9	ERNIE (Sun et al. 2019b)	91.9	92.0
QNLI	105k	5.4k	92.7	DeBERTa (He et al. 2020)	99.2	91.2
RTE	2.5k	3k	70.1	DeBERTa (He et al. 2020)	93.2	93.6
WNLI	634	146	65.1	ERNIE (Sun et al. 2019b)	95.9	95.9
ConTRoL	8.3k	804	50.6	BART-NLI-FT	61.0	94.4

Table 4: The state-of-the-art performances of popular NLI benchmarks (accuracy%).

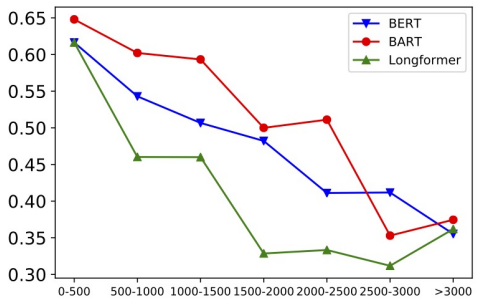

Figure 4: Performance across different context lengths.

Reasoning Type	BERT	BART
Coreferential Reasoning	74.64	74.92
Analytical Reasoning	67.96	69.65
Temporal Reasoning	56.44	57.34
Information Integration	40.07	43.39
Logical Reasoning	40.76	43.20

Table 5: Performance across reasoning types (accuracy%).

Results

Table 3 shows the main results. As shown in the table, BERT gives an overall accuracy of 50.62% and F1 of 49.49%; RoBERTa gives a higher accuracy of 45.90% and F1 of 45.67%; Longformer gives an overall accuracy of 49.88% and F1 of 46.22%; The top reported performance is given by the BART model, with a 56.34% accuracy score. Compared with human performance, the performance of BART is lower by approximately 30%. The human performance on the ConTRoL surpasses SOTA NLI models by a large margin, which demonstrate limitations for the computational models for solving contextual reasoning tasks.

As shown in Table 4, we see a huge performance drop when the SOTA model results on ConTRoL are compared to their reported score on previous NLI dataset (Liu et al. 2019). In contrast, similar to the existing benchmarks, human testees are able to achieve high scores with proper training. Different from datasets that emphasise fact extraction and verification, the inference of ConTRoL relies not only on the long-term dependency of texts, but the contextual reasoning abilities regarding long contexts.

To further understand the phenomena, we conduct various qualitative and quantitative detailed analysis on ConTRoL.

Performance Across Different Relationships

We first compare human performance and model performance across different relationships. Interestingly, as shown in Table 3, humans are good at deciding the entailment and contradiction relationship, while struggling when examining the relationship of neutral. This can be because humans tend to associate external irrelevant knowledge to the reasoning process, which is not expressed in the context. The computational models seem not to bear this burden, which gives similar results across the three labels.

Performance Across Different Context Lengths

As mentioned earlier, aside from single-paragraph context-hypothesis pairs, there are multi-paragraph context-hypothesis pairs in our dataset. We conduct experiments on the single-paragraph and multi-paragraph instances separately, which gives us the insight into how context length affects the performance of the transformer-based NLI models. The accuracy of the BERT model is 40.30% on multiple-paragraph instances while 51.17% on single-paragraph instances. We also conduct fine-grained analysis concerning the context length. The result are shown in Figure 4. When the context length increases, the model performance drops accordingly. The best model BART drops from 65% (shorter than 500 words) to 40% (longer than 3,000 words), demonstrating that ConTRoL heavily rely on passage-level reasoning ability, rather than sentence-level reasoning ability.

Performance Across Different Reasoning Types

Table 5 gives the performance over the 5 reasoning types. BERT and BART have similar trends across different reasoning types. In particular, on the coreferential reasoning type, BERT and BART give accuracies of 74.64% and 74.92%, respectively. On the other hand, both models are more confused on reasoning types including multi-step reasoning and logical reasoning. This can be because multi-step reasoning can be correlated with longer context length, and information integration is processed over multiple paragraphs. Finally performing inductive and deductive reasoning is difficult for current models, making logical reasoning a difficult endeavour (Liu et al. 2020a).

Transfer Learning

Recent studies have shown the benefit of fine-tuning on similar datasets for knowledge transfer (Huang et al. 2019). We explore three related NLI datasets for knowledge transfer, SNLI (Bowman et al. 2015), MultiNLI (Williams, Nangia,

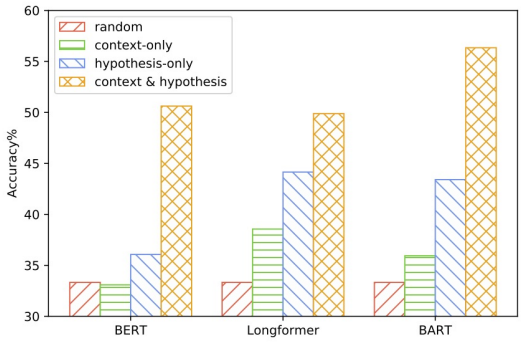

Figure 5: Ablation study on different models.

and Bowman 2018) and Adversarial NLI (Nie et al. 2020). As shown in the last two rows of Table 3, BART-NLI only achieves 45.0%, which shows that ConTRoL is different from existing NLI benchmarks. After fine-tuning on ConTRoL, BART-NLI-FT achieves the state-of-the-art results, which demonstrates that general knowledge from traditional NLI benchmarks are beneficial to the performance of ConTRoL.

Discussion

Corpus Bias

Recent studies show that pre-trained language models can make the right prediction by merely looking at the context (McCoy, Pavlick, and Linzen 2019). The hypothesis-only bias is common in large-scale datasets for NLI, particularly for benchmarks constructed by crowdsourcing methods. We conduct an ablation experiment on ConTRoL. Figure 5 shows the comparison of BERT, Longfomer and BART. BERT gives a 36.07% accuracy with hypothesis-only, which is slightly higher than theoretical random guess; Longformer gives a 44.15% accuracy, surpass BART by a small margin, which gives a 43.41% accuracy.

Context-only results are also calculated to further examine annotation artefacts in the ConTRoL dataset. BERT gives 33.09% accuracy; BART gives 35.94% accuracy; Longformer also gives a better performance than BERT and BART, which gives 38.56% accuracy. Longformer gives a better score on context-only and hypothesis-only ablation, which can be because Longformer sees more context than the other two models. The ablation results are lower than the results without ablation, which indicates that models need to look at both the contexts and the hypotheses to make the correct prediction. Thus we conclude that the ConTRoL dataset is exempt from significant annotation biases thanks to expert-designed questions.

Case Study

Figure 6 shows two cases that demonstrate the challenge in ConTRoL. P1 of Figure 6 is a representative example of the challenges brought by logical reasoning. The context concerns three athletes and three sports. We need to decide their places in a competition. The lexical overlap between the

P1: Three athletes each receive a first, second and third prize for a different sporting event. Either ***Anne*** or ***Josie*** got the second prize for Tennis. ***Anne*** got the same prize for throwing the javelin as ***Josie*** got for swimming. ***Tanya*** got the first prize for swimming, and her prize for the javelin was the same as ***Josie***'s for tennis and ***Anne***'s for swimming.
H1: *Josie was best with the javelin.*
Entailment **Contradiction ✓** **Neutral ✗**

P2: Two masked gunmen held up ***the only bank in Tuisdale*** at 10.30 on Wednesday 23 May. ***They made a successful getaway with over 500,000***. The police say that three men are helping them with their enquiries. It is also known that: Four people work at the bank. Six customers were in the bank at 10.30. No shots were fired. Ms Grainger left the bank at 10.28 on Wednesday 23 May. All the people in the bank were made to lie on the floor face down on their stomachs. The police chased the getaway car for 16 km, and then lost it. An alarm alerted the police to the hold-up. A red Ford Mondeo drove away from the bank at high speed at 10.30 on Wednesday 23 May.
H1: *As a goodwill gesture, Tuisdales other bank provided emergency access to cash for customers after their ordeal.*
Entailment **Contradiction ✓** **Neutral ✗**

Figure 6: Example mistakes of BART (✓ indicates the correct label and ✗ indicates the BART prediction. Reasoning clues are highlighted in the context.)

premise and the hypothesis is very low. BART incorrectly chooses the *Neutral* label, while we can infer from the context that Josie is actually not the best with the javelin, which can only be done by deductive reasoning. Information integration is difficult for BART.

P2 of Figure 6 shows a typical example of challenge brought by information integration, where the hypothesis is made considering the whole passage. We know from the first sentence that Tuisdale holds the only bank in the region. The hypothesis talks about the possible aftermath of the robbery, BART incorrectly chooses the *Neutral* label for it overlooks the information that Tuisdale only has one bank.In both cases, the correct answer is not explicitly mentioned in the premise, but need contextual reasoning to infer.

Conclusion

We presented the ConTRoL dataset, a passage-level NLI benchmark that consists of different contextual reasoning types. Compared with existing NLI benchmarks, the context length of the premise is bigger by a large margin, and reasoning skills such as logical reasoning, analytical reasoning and multi-step reasoning are required. Experiments show that state-of-the-art NLI models perform poorly on the ConTRoL dataset, far below human performance. Ablation study indicates that the data does not suffer from heavy annotation artefacts and can be served as a reliable NLI benchmark for future study. To our knowledge, we are the first to introduce a passage-level NLI dataset that highlights contextual reasoning.

Acknowledgements

The corresponding author is Yue Zhang. We thank the anonymous reviewers for the insightful comments. This work is supported by a grant from RxHui Inc[1].

References

Beltagy, I.; Peters, M. E.; and Cohan, A. 2020. Longformer: The Long-Document Transformer. *arXiv e-prints* .

Bhagavatula, C.; Bras, R. L.; Malaviya, C.; Sakaguchi, K.; Holtzman, A.; Rashkin, H.; Downey, D.; tau Yih, W.; and Choi, Y. 2020. Abductive Commonsense Reasoning. In *International Conference on Learning Representations*.

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 (EMNLP)*. Association for Computational Linguistics.

Clark, K.; Khandelwal, U.; Levy, O.; and Manning, C. D. 2019. What Does BERT Look at? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics.

Clark, P.; Tafjord, O.; and Richardson, K. 2020. Transformers as Soft Reasoners over Language. In Bessiere, C., ed., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, 3882–3890. International Joint Conferences on Artificial Intelligence Organization. Main track.

Cui, L.; Cheng, S.; Wu, Y.; and Zhang, Y. 2020. Does BERT Solve Commonsense Task via Commonsense Knowledge? *arXiv e-prints* arXiv:2008.03945.

Cui, L.; Wu, Y.; Liu, S.; Zhang, Y.; and Zhou, M. 2020. MuTual: A Dataset for Multi-Turn Dialogue Reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Dagan, I.; Glickman, O.; and Magnini, B. 2005. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*.

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)*. Minneapolis, Minnesota: Association for Computational Linguistics.

Giampiccolo, D.; Magnini, B.; Dagan, I.; and Dolan, B. 2007. The Third PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, RTE '07, 1–9. USA: Association for Computational Linguistics.

Giunchiglia, F. 1992. Contextual Reasoning. *EPISTEMOLOGIA, SPECIAL ISSUE ON I LINGUAGGI E LE MACCHINE* 345: 345–364.

Goldberg, Y. 2019. Assessing BERT's Syntactic Abilities. *arXiv e-prints* .

Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S. R.; and Smith, N. A. 2018. Annotation Artifacts in Natural Language Inference Data. *CoRR* abs/1803.02324.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2020. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *arXiv eprints* .

Huang, L.; Le Bras, R.; Bhagavatula, C.; and Choi, Y. 2019. Cosmos QA: Machine Reading Comprehension with Contextual Commonsense Reasoning. 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)*. Hong Kong, China: Association for Computational Linguistics.

Khot, T.; Sabharwal, A.; and Clark, P. 2018a. SciTail: A Textual Entailment Dataset from Science Question Answering. In *AAAI*.

Khot, T.; Sabharwal, A.; and Clark, P. 2018b. SciTaiL: A Textual Entailment Dataset from Science Question Answering. In *AAAI*.

Lai, A.; Bisk, Y.; and Hockenmaier, J. 2017. Natural Language Inference from Multiple Premises. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 100–109. Taipei, Taiwan: Asian Federation of Natural Language Processing.

Lai, G.; Xie, Q.; Liu, H.; Yang, Y.; and Hovy, E. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* .

Liu, J.; Cui, L.; Liu, H.; Huang, D.; Wang, Y.; and Zhang, Y. 2020a. LogiQA: A Challenge Dataset for Machine Reading Comprehension with Logical Reasoning. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* .

Liu, J.; Gardner, M.; Cohen, S. B.; and Lapata, M. 2020b. Multi-Step Inference for Reasoning Over Paragraphs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics.

[1] https://rxhui.com

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692.

McCoy, T.; Pavlick, E.; and Linzen, T. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy.

Nakhimovsky, A. 1987. Temporal Reasoning in Natural Language Understanding: The Temporal Structure of the Narrative. In *Proceedings of the Third Conference on European Chapter of the Association for Computational Linguistics*, EACL '87, 262–269. USA: 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*. Online: Association for Computational Linguistics.

Rajpurkar, P.; Jia, R.; and Liang, P. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics.

Sakaguchi, K.; Le Bras, R.; Bhagavatula, C.; and Choi, Y. 2020. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. *Proceedings of the AAAI Conference on Artificial Intelligence* .

Sap, M.; Rashkin, H.; Chen, D.; Le Bras, R.; and Choi, Y. 2019. Social IQa: Commonsense Reasoning about Social Interactions. 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)*. Hong Kong, China: Association for Computational Linguistics.

Sun, K.; Yu, D.; Chen, J.; Yu, D.; Choi, Y.; and Cardie, C. 2019a. DREAM: A Challenge Dataset and Models for Dialogue-Based Reading Comprehension. *Transactions of the Association for Computational Linguistics* .

Sun, Y.; Wang, S.; Li, Y.; Feng, S.; Chen, X.; Zhang, H.; Tian, X.; Zhu, D.; Tian, H.; and Wu, H. 2019b. Ernie: Enhanced representation through knowledge integration. *arXiv e-print* .

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. 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)*. Minneapolis, Minnesota: Association for Computational Linguistics.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *CoRR* abs/1804.07461.

Wang, H.; Yu, M.; Guo, X.; Das, R.; Xiong, W.; and Gao, T. 2019. Do Multi-hop Readers Dream of Reasoning Chains? In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. Hong Kong, China: Association for Computational Linguistics.

Welbl, J.; Stenetorp, P.; and Riedel, S. 2018. Constructing Datasets for Multi-hop Reading Comprehension Across Documents. *Transactions of the Association for Computational Linguistics* 6.

Welleck, S.; Weston, J.; Szlam, A.; and Cho, K. 2019. Dialogue Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics.

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)*. New Orleans, Louisiana: Association for Computational Linguistics.

Williams, C. C.; Kappen, M.; Hassall, C. D.; Wright, B.; and Krigolson, O. E. 2019. Thinking theta and alpha: Mechanisms of intuitive and analytical reasoning. *NeuroImage* 189: 574 – 580. ISSN 1053-8119.

Wu, Y.; Wu, W.; Xing, C.; Zhou, M.; and Li, Z. 2017. Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-Based Chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* .

Ye, D.; Lin, Y.; Du, J.; Liu, Z.; Li, P.; Sun, M.; and Liu, Z. 2020. Coreferential Reasoning Learning for Language Representation. In *Proceedings of EMNLP 2020*.

Yu, W.; Jiang, Z.; Dong, Y.; and Feng, J. 2020. ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning. In *International Conference on Learning Representations (ICLR)*.