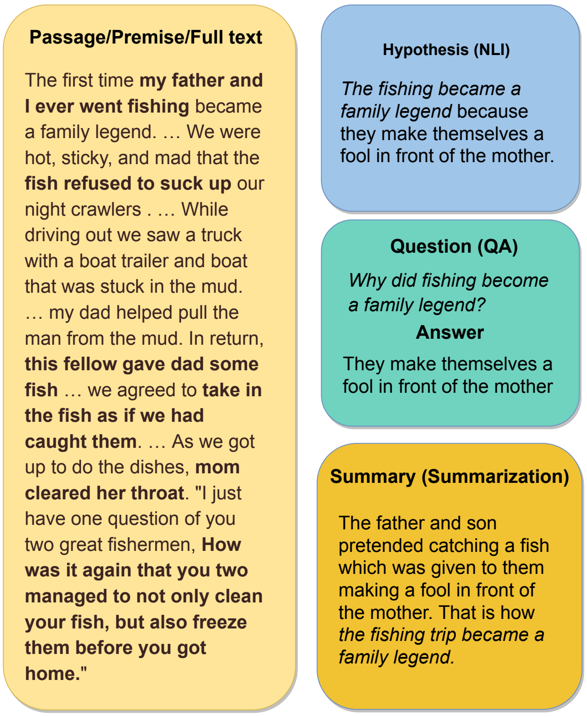
Looking Beyond Sentence-Level Natural Language Inference for Question Answering and Text Summarization

Anshuman Mishra, Dhruvesh Patel, Aparna Vijayakumar, Xiang Lorraine Li
College of Information and Computer Sciences,
University of Massachusetts Amherst

Pavan Kapanipathi, Kartik Talamadupula
IBM Research

Abstract

Natural Language Inference (NLI) has garnered significant attention in recent years; however, the promise of applying NLI breakthroughs to other downstream NLP tasks has remained unfulfilled. In this work, we use the multiple-choice reading comprehension (MCRC) and checking factual correctness of textual summarization (CFCS) tasks to investigate potential reasons for this. Our findings show that: (1) the relatively shorter length of premises in traditional NLI datasets is the primary challenge prohibiting usage in downstream applications (which do better with longer contexts); (2) this challenge can be addressed by automatically converting resource-rich reading comprehension datasets into longer-premise NLI datasets; and (3) models trained on the converted, longer-premise datasets outperform those trained using short-premise traditional NLI datasets on downstream tasks primarily due to the difference in premise lengths.



Passage/Premise/Full text

The first time **my father and I ever went fishing** became a family legend. … We were hot, sticky, and mad that the **fish refused to suck up** our night crawlers . … While driving out we saw a truck with a boat trailer and boat that was stuck in the mud. … my dad helped pull the man from the mud. In return, **this fellow gave dad some fish** … we agreed to **take in the fish as if we had caught them**. … As we got up to do the dishes, **mom cleared her throat**. "I just have one question of you two great fishermen, **How was it again that you two managed to not only clean your fish, but also freeze them before you got home.**"

Hypothesis (NLI)

The fishing became a family legend because they make themselves a fool in front of the mother.

Question (QA)

Why did fishing become a family legend?

Answer

They make themselves a fool in front of the mother

Summary (Summarization)

The father and son pretended catching a fish which was given to them making a fool in front of the mother. That is how *the fishing trip became a family legend.*

Figure 1: The tasks of Question Answering and Checking Factual Consistency of Text-Summaries can naturally be transformed into the Natural Language Inference problem.

1 Introduction

Large-scale, open Natural Language Inference (NLI) datasets (Bowman et al., 2015; Williams et al., 2018) have catalyzed the recent development of NLI models that exhibit close to human-level performance. However, the use of these NLI models for other downstream Natural Language Processing (NLP) tasks has met with limited success. Two of the most popular downstream tasks where NLI models' use has been explored are Multiple-choice Question Answering (MCRC) and Checking Factual Correctness of Summaries (CFCS) (Trivedi et al., 2019; Falke et al., 2019; Clark et al., 2018) – both of which can easily be cast into the NLI form, as shown in Figure 1. Looking closely at the composition of these datasets, it is evident that there is a stark difference in the lengths of the contexts/premises when compared to NLI datasets. As seen in Table 1, traditional NLI datasets have much shorter premises than the context texts from these downstream tasks. Prior research has shown that the capabilities required for handling local inference are very different from those required to perform inference over longer forms of text (Cooper et al., 1996; Lai et al., 2017a). In this work, we explore this conflict as a major bottleneck in the utility of NLI models (trained on traditional NLI datasets) for downstream NLP tasks. We compare the usage of long and short-premise NLI datasets on the dowsntream tasks of MCRC and CFCS, which have inherently long contexts.

Such a comparison has not been possible thus far because traditional NLI datasets do not exhibit long premises. We hence look towards recasting other tasks into NLI to generate datasets that can

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1322–1336
June 6–11, 2021. ©2021 Association for Computational Linguistics

Task	Dataset	Word Count (Avg)
NLI	SNLI	14
	Scitail	17
	MNLI	22
	RTE	42
	ANLI	54
MCRC	RACE	271
	MultiRC	252
	DREAM	110
	CosmosQA	75
CFCS	FactCC	546
	Summary Reranking	738

Table 1: The average premise (context) length in various datasets. The key point to notice here is the sharp increase in premise lengths from NLI datasets to MCRC and CFCS datasets.

be used to evaluate our conjecture. The Question-Answering (QA) task can easily be cast into the NLI form, and QA datasets (Rajpurkar et al., 2016; Lai et al., 2017b; Khashabi et al., 2018; Sun et al., 2019; Huang et al., 2019) encompass a variety of semantic phenomena that only occur in longer (con)texts. We leverage the resource-rich MCRC task to generate long-premise NLI datasets for our experiments via an automated conversion strategy.

We contrast the zero-shot model performance on the MCRC and CFCS tasks of a model pre-trained on our converted long-premise NLI dataset and a model trained on two short-premise NLI datasets - MNLI and ANLI. We show that the presence of longer premises is the primary factor for better performance on these two tasks. We further discuss other potential confounding factors for this performance difference – such as dataset vocabulary overlap and dataset conversion strategies – and eliminate the possibility of their contribution through targeted experiments.

2 Related Work

Performance on the NLI task has improved significantly due to the availability of large scale datasets (Bowman et al., 2015; Williams et al., 2018) that can be used to train data-hungry deep learning models (Kapanipathi et al., 2020; Wang and Jiang, 2015), including transformer-based architectures (Devlin et al., 2018). However, there has been very limited success in translating this performance to downstream NLP tasks. Work relevant to the use of these NLI models for downstream tasks can be categorized into two categories: (1) work focusing on using models trained on short-premise NLI datasets with fixed or learned aggregations over segmented premises to perform a target downstream task with long contexts (Falke et al., 2019; Trivedi et al., 2019); and (2) work addressing the need for task-specific NLI datasets (Kryściński et al., 2019; Demszky et al., 2018; Welleck et al., 2019).

Despite several attempts, efforts to apply models trained on available NLI datasets to downstream NLP tasks such as MCRC and CFCS have had limited success. Trivedi et al. (2019) use hand-crafted rules to first cast MCRC to NLI; and subsequently divide a long passage into smaller sentence-level premises. They use a pre-trained NLI model to evaluate per-sentence relevance scores concerning one particular hypothesis, and then combine the resulting scores using a learned representation aggregation module to assess the answer given the long passage. Falke et al. (2019) apply a similar approach for the CFCS task, and divide both the provided summary as well as the source documents into single-sentence premises and hypotheses. They use a max pooling operation over the entailment scores of all sentence-level premise-hypothesis pairs to obtain the factual correctness score for each provided summary. Both these works note that models trained on sentence-level NLI datasets do not transfer well to the MCRC and CFCS tasks. We argue that this *divide and conquer* approach is not ideal for the problem, and highlight the need for an NLI dataset with longer premises.

Another line of research focuses on re-casting datasets from other tasks into an NLI form to facilitate the direct use of NLI models on downstream tasks like MCRC and CFCS. Khot et al. (2018) use manual annotation to re-cast SciQ (a QA dataset) to SciTail – an NLI dataset. However, Clark et al. (2018) show that an NLI model trained on SciTail does not perform well on the task of MCRC. Similarly, Kryściński et al. (2019) create an automatically generated training dataset for CFCS. Even though the generated data has relatively long contexts, analysis in Zhang et al. (2020) demonstrated that a model trained on the aforementioned data showed performance improvement only when the token overlap with the source is high. Besides, Demszky et al. (2018) derive an NLI dataset by converting subsets of various QA datasets. They try two approaches for the conversion – rule-based and neural. For the rule-based approach, they extract POS tags from the question-answer pair and apply hand-crafted rules on them to convert the pair to a hypothesis sentence. Their neural approach uses a trained SEQ2SEQ BiLSTM-with-copy model (Gu et al., 2016) to convert each ⟨question, answer⟩ pair into a hypothesis sentence (the corresponding passage being the premise). While their approach looks promising, they do not show the utility of these converted datasets by training an NLI model on them. Thus, it remains unclear whether the NLI datasets generated by the conversion are beneficial for NLP tasks. We posit that this direction of research is promising and largely unexplored. In our work, we attempt to leverage the abundance of large and diverse MCRC datasets to generate long-premise NLI datasets, and show that such datasets are useful towards addressing downstream NLP tasks such as MCRC and CFCS which have inherently long contexts.

3 NLI for Downstream Tasks

Typically, NLI is cast as a multi-class classification problem, where given a premise and a hypothesis, the model classifies the relation between them as *entails*, *contradicts*, or *neutral*. For the two downstream tasks under consideration: (1) MCRC: Multiple Choice Reading Comprehension, and (2) CFCS: Checking Factual Correctness of Text-Summarization; differentiating between the *neutral* and *contradicts* class is often unnecessary. The task is thus reduced to a two-class problem; where the *contradicts* and *neutral* classes are clubbed into a *not-entails* class.

MCRC can be cast as an NLI task by viewing the given context as the premise and the transformed question-answer combinations as different hypotheses (Trivedi et al., 2019). The multiple answer-option setting can then be approached as: (a) an individual option entailment task, where more than one answer-option can be correct; or (b) a multi-class classification task across all the answer options, when only a single correct answer exists.

CFCS can also be reduced to a two-class NLI problem. A factually correct summary should be entailed by the given source text – it should not contain *hallucinated facts*, and it should also not contradict facts present in the source text.

3.1 The Long Premise Conjecture

Despite being ideally suited for reduction to NLI, both MCRC and CFCS have proved to be difficult to solve using models trained on short-premise NLI datasets (Trivedi et al., 2019; Falke et al., 2019). Datasets for these tasks contain significantly longer contexts than traditional short-premise NLI datasets (Table 1). This shift in the text length brings about a fundamental change in the nature of the NLI problem. Thus, models trained on short-premise NLI datasets are incapable of performing inference over longer texts, which we posit as the main cause for their poor performance on downstream tasks like CFCS and MCRC[*].

The paucity of manually-annotated long-premise NLI datasets poses a barrier to assessing this conjecture. We thus shift our focus towards leveraging the abundance of large and diverse MCRC datasets which can be easily recast into NLI form. While the CFCS task also provides a similar opportunity, the sheer lack of annotated training instances inhibits its use. Table 3 shows the abundance of training instances in MCRC datasets, and highlights the deficiency in CFCS datasets.

In the following section, we present our conversion strategy for reformatting MCRC datasets into long-premise NLI datasets, which are needed to test the long premise conjecture.

4 Conversion of MCRC to NLI

As shown in Figure 1, we can convert MCRC datasets into two-class NLI datasets by reusing the passage as a premise, and paraphrasing the question along with each answer option as individual hypothesis options.

We begin by using a rule-based conversion method. A dependency parse of both the question and answer option is generated using the Stanford CoreNLP package (Qi et al., 2018). This is followed by the application of conversion rules proposed by Demszky et al. (2018) to generate a hypothesis sentence. However, due to the limited coverage of rules and errors in the dependency parse, some of the generated hypotheses sound unnatural (e.g. the first example in Table 2). In order to generate more natural and diverse hypotheses and to get broader coverage in conversion, we implement a neural conversion strategy.

[*] In our experiments, we broadly consider long texts, and do not differentiate between long single sentences and multiple sentences.

	Rule-based	**Neural**	**Hybrid**
Q: What building were the four captives inside on Tuesday? **A:** CNN headquarters	The four captives inside on Tuesday were CNN headquarters.	The four captives were inside CNN headquarters on Tuesday.	The four captives were inside CNN headquarters on Tuesday.
Q: How many people were hurt when overhanging metalwork crashed onto a stage in a Toronto park Saturday afternoon. **A:** Four	Four were hurt when overhanging metalwork crashed onto a stage in a Toronto park Saturday afternoon.	Four people were hurt when overhanging metalwork crashed onto a stage in a Toronto park Saturday afternoon.. “.” # # # # “.” was the number of peo	Four were hurt when overhanging metalwork crashed onto a stage in a Toronto park Saturday afternoon.

Table 2: Examples of Rule-based, Neural and Hybrid Conversions

Due to the recent success of transformer-based text generation models, we train a BART (Lewis et al., 2019) model to generate a grammatically coherent hypothesis from question + answer option (word/phrase) as input. We use a sequence of datasets as a curriculum to finetune the BART conversion model: (1) starting with CNN/Daily Mail summarization dataset (Hermann et al., 2015), which makes the generated sentences coherent; (2) followed by Google's sentence compression dataset (Filippova and Altun, 2013), which limits the generated sequence to a single sentence; and (3) finally the annotated dataset provided by Demszky et al. (2018) which has around $71,000$ $\langle$question-answer, hypothesis$\rangle$ pairs from various QA datasets. Based on manual inspection, we find that the hypotheses generated by this method indeed sound more natural and diverse than the ones produced by the rule-based conversion[†]. In some cases, however, the generated hypotheses either discard crucial information, or contain hallucinated facts that do not convey the exact information in the source question-answer pair (Table 2). We thus define a hybrid conversion strategy, combining the desirable aspects of the rule-based and neural conversion strategies.

We design a heuristic to compose a hybrid dataset to overcome the caveats in the neural conversion. We use the number of words in the question-answer concatenation as a proxy for the expected length of the hypothesis. We target the problems of hallucination and missing information in the neural conversions by accepting only those neural-generated hypotheses that lie in the range of 0.8 and 1.2 times the length of the question-answer concatenation. We replace the rejected neural hypotheses with the rule-based hypothesis, if rule-based conversion is feasible; or with the question-answer concatenation otherwise; as seen in Table 2. The selection policy is driven by the need to get more natural and coherent conversions without compromising on the accuracy and preservation of factual information in the question and answer option. The choice of the specific range is purely empirical in nature. We use this hybrid conversion strategy to generate long-premise NLI datasets from MCRC datasets for our experiments and evaluate them in contrast to short-premise NLI datasets.

5 Experimental Setup

Our experiments involve zero-shot evaluations of pre-trained NLI models on downstream NLP tasks. In this section, we describe the transfer learning setup and the datasets used in our experiments.

5.1 A Transferable NLI model

[‡] In order to use a pretrained NLI model for MCRC and CFCS, we need that model to be agnostic to the peculiarities of the downstream task. We use a standard transfer learning setting where the model architecture is divided into two parts: (1) a transferable entailment scorer; and (2) a weight-free comparator on top of the scorer. Each premise-hypothesis pair is encoded as a single sequence, and passed through the transferable entailment scorer to produce an entailment score. Depending on the problem setup, the comparator can either be a sigmoid function (for a two-class entailment problem) as shown in Figure 2; or a softmax function (for multiple choice classification) as shown in Figure 3. This segmentation of the model makes it easy to transfer the model weights across different tasks. For the entailment scorer, we use a 2-layer feed-forward network on top of the `[CLS]` token of

[†] More examples of conversion results are presented in Appendix D.

[‡] Code available here: `https://github.com/nli-for-qa/transformers-nli`

pre-trained RoBERTa [§].

To evaluate the transferability of the entailment model, we perform various zero-shot evaluations. This requires interpreting the entailment scores a bit differently for each task. To transfer the weights from a multiple choice classification model (Figure 3) to a two class entailment model (Figure 2), we copy the weights of the transferable entailment scorer as-is, and calibrate a threshold using a dev set to interpret the outputs from the sigmoid comparator for binary classification. Since the softmax comparator does not need any calibration, the transfer in the other direction, i.e., from a two class entailment model to a multiple choice classification model is more straightforward – we simply copy the weights of the transferable entailment scorer.

5.2 Datasets

For our experiments, we use the NLI form of 4 MCRC datasets (created using the conversion method described in Section 4); 2 CFCS datasets; and 2 traditional short-premise NLI datasets. These datasets are described below:

MCRC Datasets:

RACE (Lai et al., 2017b) broadly covers detail reasoning, whole-picture reasoning, passage summarization, and attitude analysis.

MultiRC (Khashabi et al., 2018) mainly contains questions which require multi-hop reasoning and co-reference resolution.

DREAM (Sun et al., 2019) is a dialogue-based MCRC dataset, where the context is a multi-turn, multi-party dialogue.

CosmosQA (Huang et al., 2019) focuses on commonsense and inductive reasoning, which require reading between the lines.[¶]

CFCS Datasets:

FactCC (Kryściński et al., 2019) consists of tuples of the form ⟨`article, sentence`⟩, where the articles are taken from the CNN/DailyMail corpus, and sentences come from the summaries for these articles generated using several state-of-the-art abstractive summarization models.

Ranking Summaries for Correctness (evaluation set) (Falke et al., 2019) consists of articles and a set of summary alternatives for each article, where some of the provided summaries are factually inconsistent with respect to the article.

Task	Dataset	Dataset Size
MCRC	RACE	87866
	MultiRC	27243
	DREAM	6116
	CosmosQA	23766
CFCS	FactCC	931
	Summary Reranking	1000

Table 3: The number of annotated instances in MCRC and CFCS datasets. MCRC is an extremely resource-rich task whereas CFCS is considerably resource-deficient.

Short-Premise NLI Datasets:

MNLI (Williams et al., 2018) is a large-scale general domain NLI dataset that is widely used to learn and evaluate short-premise NLI models.

ANLI (Nie et al., 2019) is a large-scale NLI dataset generated through an adversarial human-in-the-loop process; where the annotations are constrained such that models trained on MNLI and SNLI predict incorrect answers. This dataset also has the longest premise lengths amongst the traditional NLI datasets compared in Table 1.

Long-Premise NLI Datasets:

We convert the following MCRC datasets to generate long-premise NLI datasets using the hybrid conversion strategy described in Section D. We refer to these datasets with a subscript *converted* attached to the source MCRC dataset.

As seen from Table 1 and Table 3, RACE is the largest dataset amongst the MCRC datasets, and also has the longest average premise length. In line with this intuition, the model trained on the $\text{RACE}_{\text{converted}}$ dataset outperforms the converted forms of other MCRC datasets (Appendix B) on all the evaluation tasks. Due to this, in the following section, we only discuss and report results on the $\text{RACE}_{\text{converted}}$ dataset for brevity and clarity of comparison. Amongst the traditional NLI datasets, we use MNLI and ANLI for a good mix of average premise lengths along with a large number of training samples.

6 Results and Discussion

Our experiments aim to answer the following questions: (1) Are long premise NLI datasets more use-

[§] The RoBERTa model is pre-trained on the masked language modeling objective as described in Liu et al. (2019). We obtain it from the HuggingFace library (Wolf et al., 2019).

[¶] Questions where the answer is "None of the above" are removed from the CosmosQA dataset.

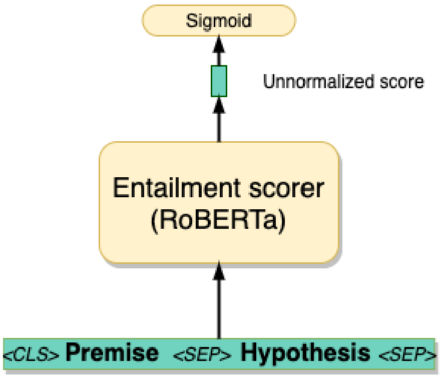

Sigmoid
Unnormalized score
Entailment scorer (RoBERTa)
<CLS> Premise <SEP> Hypothesis <SEP>

Figure 2: Two class entaiment model.

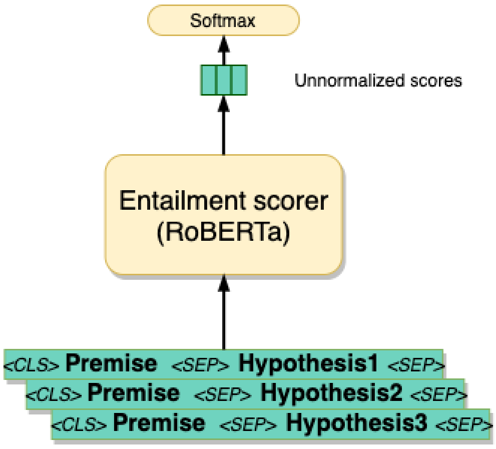

Softmax
Unnormalized scores
Entailment scorer (RoBERTa)
<CLS> Premise <SEP> Hypothesis1 <SEP>
<CLS> Premise <SEP> Hypothesis2 <SEP>
<CLS> Premise <SEP> Hypothesis3 <SEP>

Figure 3: Multiple choice classification model.

ful for downstream tasks compared to short premise NLI datasets? (Section 6.1 & 6.2); (2) How much do possible confounding factors affect our empirical evaluations? (Section 6.3). To answer these, we perform zero-shot evaluation on the MCRC and CFCS tasks.

We contrast the performance of NLI models trained on the short-premise NLI datasets (MNLI, ANLI) with one that is trained on a long-premise NLI dataset (RACE$_{\text{converted}}$). The models trained on short-premise NLI datasets are evaluated in two ways: (1) by treating the entire premise as input; and (2) by segmenting the premise into shorter segments and using a *max aggregation* over the entailment scores of all the segments (Falke et al., 2019). Since the model architecture remains the same, we use the name of the training dataset to refer to the model trained on it.

6.1 Evaluation on MCRC

For evaluating NLI models on the MCRC task, we use the hybrid conversion (Section 4) to create evaluation datasets. The MultiRC dataset contains multiple correct answer options and hence is evaluated with each question-answer option posed as a separate example. DREAM and CosmosQA datasets have only a single correct answer-option (out of 3 answer-options). Hence, for these datasets, a multi-class classification problem is posed as described in Section 3, using the model architecture described in Figure 3.

As seen in Table 4, the model trained on the long-premise RACE$_{\text{converted}}$ dataset outperforms the model trained on the short-premise NLI datasets in both regular and segmented forms of evaluation. We assert that this difference in performance can be attributed to the difference in premise lengths of the datasets. However, we allow for the possibility that using the same conversion strategy for the evaluation datasets could potentially benefit the model trained on RACE$_{\text{converted}}$. We discuss such confounding factors in Section 6.3.2.

Dataset[*] / Model	MultiRC	DREAM	CosmosQA
Random Guess	50.00	33.33	33.33
MNLI	60.58	67.76	38.11
MNLI$_{\text{segmented}}$	61.71	42.28	43.28
ANLI	67.95	74.12	49.71
ANLI$_{\text{segmented}}$	63.45	61.42	49.60
RACE$_{\text{converted}}$	**77.43**	**83.58**	**73.58**

[*] Datasets are in NLI form created using hybrid conversion method (Section 4).

Table 4: Zero-shot evaluation accuracies on MCRC datasets (in NLI form) using the transferable model architecture described in Section 5.1.

6.2 Evaluation on CFCS

Evaluations on CFCS are set up in two ways:

(1) CFCS as classification: In this form, given a document and a corresponding summary sentence, the model needs to identify if the sentence is factually correct with respect to the document (entailed) or not. In order to perform the classification, we first obtain our entailment scorer by fine-tuning the multiple choice classification model (Figure 3) on the RACE$_{\text{Converted}}$ dataset and use the dev set[‖] to calibrate a threshold[**] (described in Section 5.1) to obtain the two-class entailment model (Figure 2).

[‖] We use the dev and test dataset provided by Kryściński et al. (2019) for this task.

[**] Balanced accuracy is used to find the best threshold.

(2) CFCS as ranking: Given a source document and a set of five machine generated summaries, the model is required to rank at least one factually correct summary above all incorrect summary alternatives. Note that a variable number of these five machine generated summaries can be factually correct (Falke et al., 2019). However, there is always at least one incorrect summary in this set.

Model	Balanced Accuracy	F1-score
BERT+FactCC$_{\text{autogen}}$* #	74.15	0.51
RoBERTa	54.76	0.30
RoBERTa+MNLI	51.92	0.15
RoBERTa+MNLI$_{\text{segmented}}$	69.87	0.70
RoBERTa+ANLI	62.61	0.41
RoBERTa+ANLI$_{\text{segmented}}$	57.34	0.58
RoBERTa+RACE$_{\text{converted}}$	**86.55**	**0.73**

* These results are reported from Kryściński et al. (2019).

# FactCC$_{\text{autogen}}$ is the automatically generated training data used by Kryściński et al. (2019).

Table 5: Balanced accuracy and macro F1 score on the test set for the task of CFCS posed as a classification problem.

Model	%Correct
ESIM + SNLI *	60.70%
RoBERTa	50.47%
RoBERTa+MNLI	49.53%
RoBERTa+MNLI$_{\text{segmented}}$	66.36%
RoBERTa+ANLI	54.20%
RoBERTa+ANLI$_{\text{segmented}}$	66.35%
RoBERTa+RACE$_{\text{converted}}$	**75.70%**

* Reported from Falke et al. (2019).

Table 6: Performance of various models on the CFCS on the sentence-ranking and summary-ranking tasks. The numbers denote the fraction of highest ranked summaries which are labelled factually correct.

Table 5 and Table 6 present the results for CFCS as classification and CFCS as ranking, respectively. Similar to the MCRC task, the model trained on the long-premise RACE$_{\text{converted}}$ dataset outperforms the models trained on the short-premise NLI datasets in both regular and segmented forms of evaluation on each of the CFCS task types. Moreover, it also outperforms the FactCC model which uses the automatically generated long-premise training data (Kryściński et al., 2019).

The results of evaluations on the MCRC and CFCS tasks – which inherently contain long contexts – provide strong evidence supporting our long premise conjecture.

6.3 Confounding Factors

Natural language experiments are often vulnerable to artifacts that may leak exploitable signals into the training data that the model can fit on. Such extraneous factors, if present, can prevent the empirical isolation of the premise-length as a major factor. We therefore discuss and eliminate the two most obvious potential confounding factors.

6.3.1 Vocabulary Overlap

In the zero-shot evaluation setup, a high vocabulary overlap between the training data and the target data can potentially help a model perform better. To eliminate this confounding factor from our experiments, we calculate the vocabulary overlap of RACE, MNLI and ANLI (training data) with the 3 MCRC datasets (evaluation data). We define overlap as:

$$\frac{\text{\# words in [Vocab(train data)} \cap \text{Vocab(eval. data)]}}{\text{\# words in Vocab(eval. data)}}$$

Table 8 shows that all the datasets have similar vocabulary overlap with the three MCRC datasets. However, from Table 4, we see that the model trained on RACE$_{\text{converted}}$ considerably outperforms the models trained on the short-premise NLI datasets. This indicates that vocabulary overlap is not playing a big role in the model's performance.

To substantiate this claim, we further evaluate the two models on those subsets of the three MCRC datasets that consist only of examples where the vocabulary overlap is high (≥ 0.9). Table 7 shows that the performance of the two models on these high vocabulary overlap subsets is similar to their overall performances on the respective datasets. We can thus conclude that vocabulary overlap is not helping either of the models in terms of predictive performance.

6.3.2 Automated Conversion

We evaluate the models trained on the short-premise NLI datasets and RACE$_{\text{converted}}$ on the converted forms of the MCRC datasets. However, only the model trained on the RACE$_{\text{converted}}$ dataset is exposed to the same conversion strategy during training. It is therefore possible that the conversion

	MultiRC		DREAM		CosmosQA	
	Overall	Subset	Overall	Subset	Overall	Subset
RoBERTa+RACE$_{\text{converted}}$	77.4	77.6	83.5	85.5	73.5	74.0
RoBERTa+MNLI	60.5	61.1	67.7	68.6	38.1	37.6
RoBERTa+ANLI	67.9	68.7	74.1	73.7	49.7	49.9

Table 7: Performance of the models on high vocabulary overlap subsets of the MCRC datasets.

	MultiRC	DREAM	CosmosQA
RACE	0.905	0.974	0.852
MNLI	0.928	0.950	0.839
ANLI	0.840	0.913	0.729

Table 8: Vocabulary Overlap with MCRC datasets. The value in cell (i,j) is given by $\frac{\text{SizeOf}(\text{Vocab}_i \cap \text{Vocab}_j)}{\text{SizeOf}(\text{Vocab}_j)}$

mechanism itself becomes a confounding factor, enabling the RACE$_{\text{converted}}$ model to perform better on the MCRC task. To assess this nuance, we manually annotate a subset of the MCRC datasets using Label Studio (Tkachenko et al., 2020), with a random set of examples annotated by each of the authors. To create a setting where the difference is vivid, we design the annotation subsets such that the RACE$_{\text{converted}}$ model gives an accuracy of around 50% using the hybrid conversion strategy. The independent manual annotations prevent any exploitable signal from leaking into the training data of the model through the conversion mechanism. We compare the performance of models trained on converted forms of the RACE dataset using both our hybrid strategy as well as manual annotation.[††] We manually annotate 100 examples from MultiRC and 50 each from ComsosQA and DREAM. MultiRC is evaluated at an option-level with each question-answer pair considered an individual example. On the other hand, CosmosQA and DREAM are evaluated at a question-level, with each example consisting of three question-answer pairs, and one label corresponding to the correct answer option.

Table 9 shows that the RACE$_{\text{converted}}$ model performs better on the manually annotated subset; this eliminates the possibility of the conversion mechanism being a confounding factor in our results.

	MultiRC	DREAM	CosmosQA
Automatic	52.08	50.00	50.00
Manual	55.21	54.00	72.00

Table 9: Evaluation of the RACE$_{\text{converted}}$ model on the manually annotated subset of the MCRC datasets as compared to the same subsets with Hybrid conversion.

7 Conclusion

The difficulty of transferring entailment (NLI) knowledge to downstream NLP tasks can be largely attributed to the difference in data distributions, specifically the premise lengths. Models trained on short-premise NLI datasets are not very good at performing inference over longer texts, which is a central feature of important downstream tasks such as QA and text summarization.

We leverage the abundance of large and diverse MCRC datasets and the ease of conversion from MCRC into the NLI format to automatically and scalably create a long-premise NLI dataset to test this long-premise conjecture. We show that the long-premise nature of the converted dataset indeed helps achieve better performance on the downstream tasks of MCRC and CFCS when compared against models trained on traditional short-premise NLI datasets. We further discuss and eliminate possible confounding factors in our experiments to ensure the validity of our results.

Our work highlights a major shortcoming in popular NLI datasets that limits their usefulness to downstream NLP applications; and emphasizes the need for long-premise NLI datasets. Future work in this direction can take us closer to realizing the full potential of NLI as a fundamental task in natural language understanding.

[††] It is important to note that this setting is solely for the purpose of establishing the role of the hybrid conversion strategy as a potential confounding factor in the performance of the RACE$_{\text{converted}}$ model. The absolute accuracy numbers are not reflective of the model performance on the overall dataset.

Ethical Considerations

In this work, we use open source datasets, libraries, and services which are freely available and appropriately cited. We do not release the converted form of the MCRC dataset in respect of existing copyright; however, we provide all the information required to reproduce our experimental setup, datasets, and results in the content of the main paper as well as in the appendix.

All rules used in the conversion strategies (Section 4), as well as the manual annotations performed as part of the confounding factors analysis, were produced solely by the group of authors. Our work did not involve any external human subjects; and did not require institutional review.

Looking forward, it is certainly possible that the neural conversion strategy proposed by us in Section 4 may be applied by readers of this work in other – potentially scaled-up – contexts. Since the conversion is used as a means to an end (producing an appropriate long-premise dataset) rather than as the central contribution of the current work, we do not provide an extended analysis of the pros and cons of this strategy.

Acknowledgment

We thank Rajarshi Das, and Andrew McCallum for their constructive feedbacks in the 696DS class during Spring 2020. We also thank the anonymous reviewers for their insightful suggestions.

Xiang Lorraine Li is supported in part by IBM Research AI through the AI Horizons Network and in part under the University of Southern California subcontract no. 123875727 under Office of Naval Research prime contract no. N660011924032. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, et al. 1996. Using the framework, fracas: a framework for computational semantics. Technical report, Technical Report LRE 62-051 D-16, The FraCaS Consortium.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. *ArXiv*, abs/1809.02922.

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

Tobias Falke, Leonardo FR Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220.

Katja Filippova and Yasemin Altun. 2013. Overcoming the lack of parallel data in sentence compression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1481–1491, Seattle, Washington, USA. Association for Computational Linguistics.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1693–1701. Curran Associates, Inc.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. *arXiv preprint arXiv:1909.00277*.

Pavan Kapanipathi, Veronika Thost, Siva Sankalp Patel, Spencer Whitehead, Ibrahim Abdelaziz, Avinash Balakrishnan, Maria Chang, Kshitij Fadnis, Chulaka Gunasekara, Bassem Makni, Nicholas Mattei, Kartik Talamadupula, and Achille Fokoue. 2020. Infusing knowledge into the textual entailment task using graph convolutional networks. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

Alice Lai, Yonatan Bisk, and Julia Hockenmaier. 2017a. Natural language inference from multiple premises. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 100–109, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. 2017b. Race: Large-scale reading comprehension dataset from examinations. In *EMNLP*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. *arXiv preprint arXiv:1804.08207*.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*.

Kyle Richardson, Hai Hu, Lawrence S Moss, and Ashish Sabharwal. 2019. Probing natural language inference models through semantic fragments. *arXiv preprint arXiv:1909.07521*.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. Dream: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Maxim Tkachenko, Mikhail Malyuk, Nikita Shevchenko, Andrey Holmanyuk, and Nikolai Liubimov. 2020. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Harsh Trivedi, Heeyoung Kwon, Tushar Khot, Ashish Sabharwal, and Niranjan Balasubramanian. 2019. Repurposing entailment for multi-hop question answering tasks. 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)*, pages 2948–2958.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228*.

Shuohang Wang and Jing Jiang. 2015. Learning natural language inference with lstm. *arXiv preprint arXiv:1512.08849*.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *ACL*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 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)*, pages 1112–1122. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Yuhui Zhang, Yuhao Zhang, and Christopher D. Manning. 2020. A close examination of factual correctness evaluation in abstractive summarization.

A Quality of Converted Datasets

We evaluate the quality of the converted datasets by using benchmarks that probe the trained models for semantic phenomenon (Poliak et al., 2018; Richardson et al., 2019). Tables 10 and 11 shows the performance of the models on the different semantic phenomena. We show that the converted NLI datasets are at par, and sometimes better than a specially curated NLI dataset such as MNLI. For the purpose of illustration, we report results on the $\text{RACE}_{\text{converted}}$ dataset and MNLI.

B Comparison of Converted MCRC Datasets

We use the hybrid conversion strategy discussed in Section 4 to generate long-premise NLI datasets from each of the MCRC datasets – RACE, MultiRC, DREAM and CosmosQA. As seen from Table 1 and 3, RACE has the longest average premise length as well as the most number of training examples. It is thus, intuitive to see from Tables 12, 13 and 14 that the $\text{RACE}_{\text{converted}}$ model outperforms the other converted models in each of the tasks.

C Reproducibility Checklist

C.1 Details of the datasets used

Table 15 gives the train/dev/test splits of the various source datasets used in this work. We follow the same splits after the conversion to NLI form. Since the test datasets are not openly available for MultiRC and CosmosQA, we use the corresponding dev sets to report our results.

Table 16 shows the proportion (absolute numbers) of neural, rule-based and Q+A examples in the final hybrid datasets.

C.2 Neural Conversion

We use the following training sequence to obtain the final neural conversion model:

1. Obtain the pre-trained BART model (Lewis et al., 2019) fine-tuned on CNN/Dailymail from HuggingFace library.[*]

2. Fine-tune the model using the hyperparameters mentioned in Table 17 on google-sentence completion dataset (Filippova and Altun, 2013)[†]

3. Further fine-tune the model on the QA2D datatset (Demszky et al., 2018).[‡]

C.2.1 Experiments

- The hyperparams for the models used throughout the Section 6 are shown in Table 18. These were obtained using minimal manual tuning.

- The threshold for CFCS as classification experiments (Section 6.2 (1)) we calculated by tuning for best balanced accuaracy https://scikit-learn.org/stable/modules/generated/sklearn.metrics.balanced_accuracy_score.html.

D Conversion examples

Tables 19, 20 and 21 show examples of rule-based and neural conversions on RACE, MultiRC and DREAM respectively.

[*] https://huggingface.co/facebook/bart-large-cnn

[†] https://github.com/google-research-datasets/sentence-compression

[‡] https://worksheets.codalab.org/worksheets/0xd4ebc52cebb84130a07cbfe81597aaf0/

Semantic Fragment	**RACE$_{\text{converted}}$**	**MNLI**
Boolean	84	93.0
Comparative	66.8	74.5
Conditional	83.4	61.1
Counting	45.3	54.0
Monotonicity (Hard)	50.9	62.8
Monotonicity (Simple)	67.6	62.0
Negation	86.8	82.5
Quantifier	71	77.6
Overall	69.48	70.94

Table 10: Results on the Semantic Fragments dataset from (Richardson et al., 2019).

Sem. Phenomenon	**RACE$_{\text{converted}}$**	**MNLI**
Factuality	61.47	**67.39**
NER	**69.28**	58.36
Pun	49.71	**58.54**
Sentiment	**94.17**	68.5
Lexicosyntactic (MV)	35.62	**43.49**
Lexicosyntactic (VN)	48.13	**55**
Lexicosyntactic (VC)	**50.25**	46.59
Overall	**63.22**	56.08

Table 11: Results on the Diverse Natural Language Inference dataset from (Poliak et al., 2018). RACE$_{\text{Neural}}$ refers to the converted RACE dataset created using the BART model. MV refers to the MegaVeridicality dataset; VN to the VerbNet dataset; VC to the VerbCorner dataset.

Dataset[*] **Model**	RACE (271)	MultiRC (252)	DREAM (110)	CosmosQA (75)
Random Guess	25.00	50.00	33.33	33.33
MultiNLI	44.34	60.58	67.76	38.11
MultiNLI$_{\text{Segmented}}$	41.01	61.71	42.28	43.28
RACE$_{\text{converted}}$	*83.99*	**77.43**	**83.58**	**73.58**
MultiRC$_{\text{converted}}$	58.02	*81.22*	67.12	43.65
DREAM$_{\text{converted}}$	**65.01**	71.08	*83.99*	61.00
CosmosQA$_{\text{converted}}$	49.27	48.80	72.46	*83.89*

[*] Datasets are in NLI form created using hybrid conversion method (Section **??**) for the models trained on the converted datasets.

Table 12: Zero-shot evaluation accuracies achieved by models trained on converted NLI datasets and MultiNLI on *other* MCRC datasets (in NLI form) using the transferable model architecture described in Section 5.1. The numbers in the parenthesis of the column headers denote the average premise lengths of the datasets.

Model	**Balanced Accuracy**	**F1-score**
BERT+FactCC$_{\text{autogen}}$[*] [#]	74.15	0.51
RoBERTa	54.76	0.30
RoBERTa+MultiNLI	51.92	0.15
RoBERTa+MultiNLI$_{\text{segmented}}$	69.87	0.70
RoBERTa+CosmosQA$_{\text{converted}}$	55.96	0.52
RoBERTa+DREAM$_{\text{converted}}$	75.69	0.69
RoBERTa+MultiRC$_{\text{converted}}$	82.03	0.72
RoBERTa+RACE$_{\text{converted}}$	**86.55**	**0.73**

[*] These results are reported from Kryściński et al. (2019).
[#] FactCC$_{\text{autogen}}$ is the automatically generated training data used by Kryściński et al. (2019).

Table 13: Balanced accuracy and macro F1 score on the test set for the task of CFCS posed as a classification problem.

Model	**% Correct**	
	Sentence-pair Ranking	**Summary Ranking**
ESIM + SNLI [*]	67.60%	60.70%
BERT+FactCC$_{\text{autogen}}$ [†] [#]	70.00%	-
QAGS[‡]	72.10%	-
RoBERTa	56.03%	50.47%
RoBERTa+MultiNLI	81.76%	49.53%
RoBERTa+MultiNLI$_{\text{segmented}}$	81.23%	66.36%
RoBERTa+CosmosQA$_{\text{converted}}$	76.41%	49.53%
RoBERTa+DREAM$_{\text{converted}}$	78.28%	68.22%
RoBERTa+MultiRC$_{\text{converted}}$	72.21%	67.23%
RoBERTa+RACE$_{\text{converted}}$	**86.59%**	**75.70%**

[*] [†] [‡] Reported from Falke et al. (2019), Kryściński et al. (2019) and Wang et al. (2020), respectively.
[#] FactCC$_{\text{autogen}}$ is the automatically generated training data for their model.

Table 14: Performance of various models on the CFCS on the sentence-ranking and summary-ranking tasks. The numbers denote the fraction of highest ranked summaries which are labelled factually correct.

Dataset	Number of examples		
	Train	Dev	Test
RACE	87866	4887	4934
MultiRC	27243	4848	-
DREAM	6116	2040	2041
CosmosQA	6116	2040	-
FactCC	-	931	503
Sentence Ranking	-	746	-
Summary Ranking	-	2555	530

Table 15: Number of examples in each of the datasets.

Dataset	**Split**	**Neural**	**Rule-based**	**Q+A**
RACE	Train	314448	16808	20208
	Dev	17447	912	1189
	Test	18284	580	872
MultiRC	Train	23613	3630	0
	Dev	4156	692	0
DREAM	Train	16708	1530	110
	Dev	5531	531	58
	Test	5588	495	40
CosmosQA	Train	7298	848	32
	Dev	60009	10889	400

Table 16: The proportion (absolute numbers) of neural, rule-based and Q+A examples in the hybrid datasets.

Hyperparam	**Dataset/fine-tune curriculum step**	
	Google-sentence compression	QA2D
learning rate	1e-5	1e-5
weight decay	0.01	0.01
adam epsilon	1e-8	1e-8
max. grad. norm	1.0	1.0
warmup steps	1125	600
batch size	24	32
max epochs	3	5
max seq. len	50	50
lower-case	False	False
Runtime metrics		
Python	3.7.4	3.7.4
GPU Type	GeForce RTX 2080 Ti	GeForce RTX 2080 Ti
Num. GPUs	1	1

Table 17: Hyperparameters and runtime metrics for training the neural conversion model

	Model				
Hyperparam	**RoBERTa+RACE**	**RoBERTa+DREAM**	**RoBERTa+MultiRC**	**RoBERTa+CosmosQA**	**RoBERTa+MultiNLI**
learning rate	1e-5	1e-5	1e-5	1e-5	1e-5
weight decay	0.001	0.1	0.001	0.1	0.01
max. grad. norm.	1.0	1.0	1.0	1.0	1.0
warmup steps	1300	500	300	500	1200
batch size	24	32	32	24	48
max epochs	4	10	4	4	4
Runtime metrics					
Python	3.7.3	3.7.3	3.7.3	3.7.3	3.7.3
GPU type	m40	m40	m40	m40	Titan X
Num. GPUs	1	1	1	1	1

Final dev accuracy	83.08 (Q+A)
	82.02(Neural)
	84.00(Hybrid)
		84.36 (Q+A)
		84.07 (Neural)
		84.12 (Hybrid)
			84.28 (Q+A)
			80.16(Neural)
			79.94 (Hybrid)
				85.33 (Q+A)
				83.65 (Neural)
				83.91 (Hybrid)
					93.44

Table 18: Hyperparam setting for the models trained on MCRC datasets and MultiNLI (same for Q+A, Neural, and Hybrid from). These are common for all models in the experiments (Section 6).

	Rule-based	**Neural**
Q: How do suburban commuters travel to and from the city in Copenhagen at present? **A:** About one third of the suburban commuters travel by bike.	Suburban commuters travel to about one third of the suburban commuters travel by bike and from the city in Copenhagen at present.	Suburban commuters travel to and from the city in Copenhagen at present by bike
Q: What's the best title of the passage? **A:** Blame! Blame! Blame!	The best title of the passage's blame.	The best title of the passage is Blame! Blame! blame! blamage!
Q: What influence did the experiment have on Alexander ? **A:** He realized that slowing down his life speed could bring him more content.	The experiment had he realized that slowing down his life speed could bring him more content on Alexander.	The experiment influenced Alexander to realize that slowing down his life speed could bring him more content.
Q: Which of the following is TRUE about the report findings? **A:** The reading scores among older children have improved.	The reading scores among older children have improved is TRUE.	It is true that the reading scores among older children have improved.

Table 19: Examples of Rule-based and Neural Conversions on RACE.

	Rule-based	**Neural**
Q: Timothy likes to spend his time after school doing what and with who? **A:** Timothy likes to play sports.	Timothy likes to spend his time after school doing what and with Timothy likes to play sports.	Timothy likes to play sports after school.
Q: What building were the four captives inside on Tuesday? **A:** CNN headquarters	The four captives inside on Tuesday were CNN headquarters.	The four captives were inside CNN headquarters on Tuesday.
Q: How might Air New Zealand's video partner benefited from helping to make this video? **A:** Coincides with the 50th anniversary of Sports Illustrated's Swimsuit franchise	Air New Zealand's video partner might benefited from helping to make this video by coincides with the 50th anniversary of Sports Illustrated's Swimsuit franchise.	Air New Zealand's video partner benefited from helping to make this video because it coincides with the 50th anniversary of Sports Illustrated's Swimsuit franchise.
Q: Did Alexander set out to secure his northern fronts and was he able to accomplish this goal? **A:** Yes and yes.	⟨ Unable to Convert ⟩	Alexander set out to secure his northern fronts and was he able to accomplish this goal.

Table 20: Examples of Rule-based and Neural Conversions on MultiRC

	Rule-based	**Neural**
Q: What is one method of treatment the dentist does NOT mention? **A:** doing a root canal	Doing a root canal is one method of treatment the dentist NOT mentions.	One method of treatment the dentist does NOT mention is doing a root canal.
Q: How often does the woman see her parents? **A:** Once a week.	The woman sees her parents once a week.	The woman sees her parents once a week.
Q: What does the man think of the woman's idea at first? **A:** He strongly opposes it.	The man thinks he strongly opposes it of the woman's idea at first.	The man strongly opposes the woman's idea at first.
Q: What does the man think of the teacher? **A:** She's from Asia.	The man thinks she's from Asia of the teacher.	The man thinks the teacher is from Asia.

Table 21: Examples of Rule-based and Neural Conversions on DREAM