Looking Beyond Sentence-Level Natural Language Inference for Downstream Tasks

Anshuman Mishra [*1], Dhruvesh Patel[*1], Aparna Vijayakumar[*1],
Xiang Li[1], Pavan Kapanipathi[2] and Kartik Talamadupula[2]

[1] *College of Information and Computer Sciences, University of Massachusetts Amherst*
[2]*IBM Research*

Abstract

In recent years, the Natural Language Inference (NLI) task has garnered significant attention, with new datasets and models achieving near human-level performance on it. However, the full promise of NLI – particularly that it learns knowledge that should be generalizable to other downstream NLP tasks – has not been realized. In this paper, we study this unfulfilled promise from the lens of two downstream tasks: question answering (QA), and text summarization. We conjecture that a key difference between the NLI datasets and these downstream tasks concerns the length of the premise; and that creating new long premise NLI datasets out of existing QA datasets is a promising avenue for training a truly generalizable NLI model. We validate our conjecture by showing competitive results on the task of QA and obtaining the best reported results on the task of Checking Factual Correctness of Summaries.

1 Introduction

Natural Language Inference (NLI) is the task of determining the relation between a given premise-hypothesis text pair; and is critical for natural language understanding. The availability of large-scale, open NLI datasets (Bowman et al., 2015; Williams et al., 2018) has recently resulted in the development of bigger and more robust models for solving the task of NLI. As some of these models close in on human-level performance, a natural question arises: *can models trained on these large-scale NLI datasets be used for other downstream NLP tasks?* So far, efforts towards using NLI for downstream tasks have had limited success (Trivedi et al., 2019; Falke et al., 2019; Clark et al., 2018).

One potential reason for this lack of success may be the inherent nature of the existing NLI

---

*Equal contribution.

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

datasets. Specifically, most existing NLI datasets consider one or at most a few sentences as the premise; and hence, can be tackled successfully by models that possess an understanding of only *local* sentence-specific semantics (negation, quantification, conditionals, monotonicity, etc.). On the other hand, most downstream NLP tasks of interest such as Question Answering (QA) and Text-Summarization require reasoning over much longer texts. While it has been posited that the capabilities required for handling sentence-level inference are very different from those required to perform inference on longer forms of text (Cooper et al., 1996; Lai et al., 2017a), the effect of this on downstream

tasks has not been studied.

In order to investigate this, we need to evaluate existing models – trained on sentence-level NLI datasets – on datasets that feature NLI instances with longer premises. However, current NLI datasets do not exhibit long premises. QA datasets (Rajpurkar et al., 2016; Lai et al., 2017b; Khashabi et al., 2018; Sun et al., 2019; Huang et al., 2019) on the other hand, encompass a variety of multi-sentence semantic phenomena. We thus work towards transforming these QA datasets into NLI datasets with long premises. We evaluate models trained on these transformed datasets on two downstream tasks - Multiple-Choice Reading Comprehensions (MCRC) in the QA domain and Checking Factual Correctness of Summaries (CFCS) in the text-summarization domain. Both of these tasks can be reduced to an NLI form (Figure 1).

The main contributions of this paper are as follows: (1) We argue that models trained on existing NLI datasets lack the multi-sentence reasoning capabilities that are needed for downstream tasks such as Question Answering and Summarization. (2) To train NLI models capable of multi-sentence reasoning, we present and analyze three different conversion methods to transform existing MCRC datasets to multi-sentence NLI datasets. We validate the quality of the converted datasets by showing that models trained on them have performance competitive to existing MCRC models. (3) Our results and analysis show that due to the presence of multi-sentence premises, models trained on the converted NLI datsets perform better than those trained on single-sentence NLI datasets, on both MCRC and CFCS downstream tasks.

2 Related Work

NLI has gained significant attention 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, work relevant to the use of these NLI models for downstream tasks has been very limited and can be categorized into two categories: (1) work focusing on using models trained on sentence-level NLI datasets with fixed or learned aggregation to perform a target downstream task (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).

Recent efforts to apply models trained on sentence-level NLI datasets on downstream NLP tasks such as MCRC and CFCS have had limited success. Trivedi et al. (2019) use simple rules to first cast the problem of MCRC to NLI. and subsequently divide the long passage into smaller sentence-level premises. They use a pretrained NLI model to obtain sentence-level relevance scores with respect to a particular hypothesis combined with a *learned* representation aggregation module to obtain the score for that hypothesis. Falke et al. (2019) apply a similar approach for the task of CFCS, and divide both the provided summary as well as the source documents into single-sentence premises and hypotheses. They use a simple entailment score aggregation over 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 task of MCRC and CFCS. We argue that this *divide and conquer* approach is not ideal for the problem, and highlight the need for a *native* multi-sentence inference model.

To facilitate the direct use of NLI models on downstream tasks like MCRC and CFCS, an interesting alternate approach has been to re-cast datasets from other tasks into NLI datasets. 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 this data has premises consisting of multiple sentences, the analysis done by Zhang et al. (2020) finds that a model trained on this data works well only for summaries having high token overlap with the source. Demszky et al. (2018) attempt to create an NLI dataset that requires inter-sentence reasoning by converting subsets of various QA datasets to NLI. They try two approaches for the conversion – rule-based and neural. 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. This makes it unclear whether the NLI datasets generated by the conversion are useful for any downstream task. We posit that this direction of research is promising and largely unexplored. Hence, in our work, we attempt to leverage the broad spectrum of MCRC datasets by recasting them to NLI datasets, and show their usefulness by performing the downstream tasks of MCRC and CFCS.

3 NLI for Downstream Tasks

NLI is usually 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*. It can also be cast into a two-class problem, where the *contradicts* and *neutral* classes are clubbed into a *not-entails* class. For all our experiments and analysis, we pose NLI as a two-class problem. We investigate the usefulness of NLI for the downstream tasks of Multiple Choice Reading Comprehension (MCRC) and Checking Factual Correctness of Text-Summarization (CFCS).

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) individual option entailment tasks, where more than one answer-option can be correct; or b) a multi-way classification task by selecting the answer-option which gets the highest entailment score from the model, when only a single correct answer-option 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.

Despite being ideally suited for reduction to NLI, both MCRC and CFCS have proved to be difficult to solve using models trained on single-sentence NLI datasets (Trivedi et al., 2019; Falke et al., 2019).

3.1 The Long Premise Conjecture

Datasets for the downstream MCRC and CFCS tasks contain significantly longer texts than the single-sentence NLI datasets (Table 1). This shift in the text length brings about a fundamental change

Task	Dataset	Word Count (Avg)	Sentence Count (Avg)
NLI	MultiNLI	22	1.1
	SNLI	14	1.0
MCRC	RACE	271	18.5
	MultiRC	252	14.3
	DREAM	110	13.9
	CosmosQA	75	3.8
CFCS	FactCC	546	28.5
	Summary Reranking	738	29.5

Table 1: The average premise 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.

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

Table 2: The number of annotated instances in MCRC and CFCS datasets. MCRC is an extremely resource-rich task whereas CFCS is considerably resource-deficient.

in the nature of the NLI problem. Performing inference over longer forms of text requires a multitude of additional reasoning skills like coreference resolution, event detection, dialogue understanding, abductive reasoning etc. (Cooper et al., 1996; Lai et al., 2017a; Demszky et al., 2018). These are over and above the reasoning types needed to perform inference locally at sentence level. Thus, models trained on sentence-level NLI datasets are incapable of performing multi-sentence inference, which we posit as the main cause for their low performance on downstream tasks like CFCS and MCRC.

In order to train models capable of performing multi-sentence inference, we need NLI datasets that possess longer multi-sentence premises. The challenge, however, is to obtain such datasets. The paucity of multi-sentence NLI datasets can be overcome by transforming large MCRC datasets into NLI datasets through a quality preserving transformation procedure. While the task of CFCS also

provides a similar opportunity, the sheer lack of annotated training instances inhibits its use. Table 2 shows the abundance of training instances in MCRC datasets, and highlights the deficiency in CFCS datasets. Hence, in this work, we rely on various MCRC datasets to provide this data.

In the following section, we present three methods to reformat MCRC datasets to create multi-sentence NLI datasets. We then evaluate models trained on these multi-sentence NLI datasets on the tasks of MCRC and CFCS, and contrast their performance with those trained on a single-sentence NLI dataset.

4 Reformatting 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. The following describes the different conversion techniques we use for this.

4.1 Rule-based Conversion

In the rule-based method of conversion, we use the Stanford CoreNLP package (Qi et al., 2018) to generate the dependency parse of both the question and the answer option, 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 (first example in Table 3). In order to generate more natural and diverse hypotheses and to get broader coverage in conversion, we implement a neural conversion strategy.

4.2 Neural Conversion

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 71000 (question-answer, hypothesis) 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 some crucial information, or contain hallucinated facts that do not convey the exact information in the source question-answer pair (Table 3). We thus define a hybrid conversion strategy, combining the desirable aspects of the rule-based conversion and the neural conversion strategies.

4.3 Hybrid Conversion

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 3. 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. In Section 7, we discuss in detail the effectiveness of the heuristically combined dataset.

5 A Transferable NLI model

In order to use pretrained NLI models for the tasks of MCRC and CFCS, we need the model to be agnostic to the peculiarities of the downstream task. This can be achieved by dividing the model architecture into two parts : (1) a transferable entailment scorer; and (2) a weight-free comparator on top of that scorer. Each premise-hypothesis pair is encoded as a single sequence pair 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

---

*We refer to the annotated dataset provided by Demszky et al. (2018) as QA2D.

†More examples of conversion results are presented in Appendix B.

	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 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	Suburban commuters travel to about one third of the suburban commuters travel by bike and from the city in Copenhagen at present.

Table 3: Examples of Rule-based, Neural and Hybrid Conversions

(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 logical 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 pretrained RoBERTa [‡].

In our experiments evaluating 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.

6 Datasets

For our experiments, we use 4 MCRC datasets and their transformed NLI versions; 2 CFCS datasets; and 1 single-sentence NLI dataset. These datasets are qualitatively described below:

Single-sentence NLI Dataset:

MultiNLI (Williams et al., 2018) is chosen as the single-sentence NLI dataset as it is widely used to learn and evaluate sentence-level NLI models.

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 coreference 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 more 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 w.r.t the article.

7 Experiments and Results

In this section, we discuss the quality of the converted datasets, and the ability of models trained on these datasets to transfer knowledge to the downstream tasks of MCRC and CFCS. We contrast the performance of these models with a model trained on MultiNLI to assert the utility of the converted datasets on long premise downstream tasks; and in this process evaluate the long premise conjecture.

[‡]The RoBERTa model is pretrained on the masked language modelling objective as described in Liu et al. (2019). We obtain it from the HuggingFace library (Wolf et al., 2019).

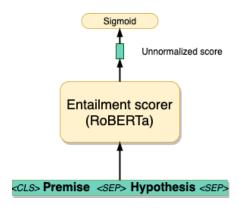[§]Questions where the answer is "None of the above" are removed from the CosmosQA dataset.

Figure 2: Two class entaiment model.

Figure 3: Multiple choice classification model.

7.1 Evaluating Conversion Quality

In order to evaluate the quality of conversion, we compare the NLI models trained on the converted datasets to their corresponding MCRC QA models. For this, we finetune RoBERTa in the multiple-choice classification setting (Figure 3) on each of the converted datasets. In order to set the performance bar, we also train RoBERTa Q+A concatenation models[¶] on each of the corresponding MCRC datasets.

Dataset	Dataset Format (conversion method)		
	QA (Q+A)	NLI (Neural)	NLI (Hybrid)
RACE	84.33	82.89	83.99
DREAM	84.22	82.41	83.29
MultiRC	85.19	80.60	81.22
CosmosQA	85.58	83.34	83.89

Table 4: Test set accuracy of models trained on converted forms of different MCRC datasets, formed using the three conversion strategies described in Section 4.

Table 4 shows that models trained on the converted datasets achieve performance comparable to the corresponding Q+A models for each of the four MCRC datasets. From this result, we can infer that the conversion mechanism captures most of the information from the MCRC datasets. Further, the models trained on the datasets formed by the hybrid conversion strategy perform consistently better in comparison to their pure neural counterparts. This
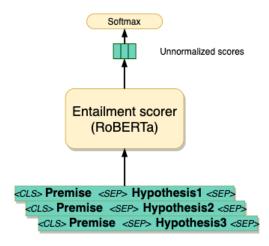

[¶]Q+A concatenation form follows the work of Liu et al. (2019) and sets a very strong quality bar.

substantiates the motivation for performing the hybrid conversion strategy as discussed in Section 4.3, and shows that the resulting hybrid conversion approach does produce better quality conversions than neural or rule-based alone. Hence, we only use the NLI dataset obtained using the hybrid conversion technique for our analysis in the subsequent experiments.

Having ascertained the quality of the converted NLI datasets, we discuss the experiments performed to substantiate our long premise conjecture by performing the task of MCRC and CFCS.

7.2 Long Premise Conjecture

To validate the long premise conjecture, we perform the tasks of MCRC and CFCS using pretrained NLI models. Specifically, we analyze and contrast the performance of NLI models trained on the sentence-level NLI dataset - MultiNLI, with those trained on multi-sentence NLI datasets obtained by converting the four MCRC datasets using the hybrid conversion strategy described in Section 4. We compare the zero-shot performance of these models on the MCRC and CFCS datasets described in Section 6. All MCRC evaluations are performed using the transformed NLI version of the data. Since MultiNLI is a single-sentence NLI dataset, the model trained on MultiNLI is evaluated in two ways: (1) by passing the entire premise; and (2) by segmenting the premise into individual sentences and aggregating the entailment score with respect to all the segments.

7.2.1 Evaluation on MCRC

We evaluate each of the five models on all the MCRC datasets (in NLI form) and discuss the performances here. The results of these evaluations

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$_{Segmented}$	41.01	61.71	42.28	43.28
RACE$_{Hybrid}$	x	**77.43**	**83.58**	**73.58**
MultiRC$_{Hybrid}$	58.02	x	67.12	43.65
DREAM$_{Hybrid}$	**65.01**	71.08	x	61.00
CosmosQA$_{Hybrid}$	49.27	48.80	72.46	x

[†] Datasets are in NLI form created using hybrid conversion method (Section 4.3) for the hybrid models

[x] These numbers are not presented as they are not the result of zero-shot evaluation. Refer Table 4 for them.

Table 5: 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. The numbers in the parenthesis of the column headers denote the average premise lengths of the datasets.

Figure 4: The graphs show the performance of models trained on RACE$_{Hyrbid}$ and MultiNLI at different premise lengths on a combined evaluation set of all the MCRC datasets mentioned in Section 6. . The accuracy at length x denotes the accuracy of the models on the examples with premise length in $[x, x + 50)$ words.

are presented in Table 5. We show that, in most cases, the models trained on the converted NLI datasets outperform the MultiNLI model on all *other* [‖] MCRC datasets. We assert that this difference in performance can be attributed to the difference in premise lengths of the converted datasets and MultiNLI.

To present further evidence in support of our claim, we analyse the performance of the models trained on the hybrid conversion of RACE (RACE$_{Hybrid}$) and MultiNLI, with varying premise length. For the purpose of this experiment, we combine all the MCRC dev datasets described in Section 6 into a single large dev set, and plot the performance with respect to the number of words in the premise in buckets of size 50. Figure 4 shows the sharp decline in the performance of the MultiNLI model as the length of the premise increases beyond 150 words, whereas RACE$_{Hybrid}$ is much more robust to increases in premise length.

7.2.2 Evaluation on CFCS

We show the utility of long-premise NLI datasets by performing the CFCS task, which can be set up in the following 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 (is entailed) or not. In order to perform the classification, we first obtain our entailment
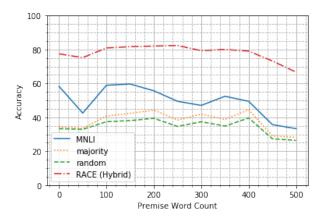
scorer by fine-tuning the multiple choice classification model (Figure 3) on the NLI form of the RACE dataset and use the dev set[**] to calibrate a threshold[††] as described in Section 5 to obtain the two-class entailment model (Figure 2).

(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. We also solve the auxiliary diagnostic task of sentence-pair ranking, where the premise document $D = \{d_1, d_2, \ldots\}$ as well as the summary $S = \{s_1, s_2, \ldots\}$ are divided into individual sentences and the model is required to decide if d_i entails s_j or not.

Table 6 and Table 7 present the results for CFCS as classification and CFCS as ranking, respectively. As can be seen, the model performances steadily increase as the premise lengths in the training data increase. The model trained on RACE$_{Hybrid}$ which has the longest average premise length (c.f. Table 1), outperforms all the models trained on datasets having comparatively shorter premises. Moreover, it also outperforms the FactCC model which uses the automatically generated long-premise training data (Kryściński et al., 2019). Another insightful

---

[‖] Other refers to all the MCRC datasets (shown in Section 6) except the one on which the model is trained.

[**] 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.

[‡‡] A variable number of these five machine generated summaries can be factually correct. However, there is always at least one incorrect summary in this set.

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

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

Table 6: 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$_{autogen}$ † #	70.00%	-
QAGS‡	72.10%	-
RoBERTa	56.03%	50.47%
RoBERTa+MultiNLI	81.76%	49.53%
RoBERTa+MultiNLI$_{Segmented}$	81.23%	66.36%
RoBERTa+CosmosQA$_{Hyrbid}$	76.41%	49.53%
RoBERTa+DREAM$_{Hyrbid}$	78.28%	68.22%
RoBERTa+MultiRC$_{Hyrbid}$	72.21%	67.23%
RoBERTa+RACE$_{Hyrbid}$	**86.59%**	**75.70%**

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

Table 7: 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.

observation is that the model trained on MultiNLI performs the short-premise task of sentence-pair ranking reasonably well, but is unable to translate this to the task of summary ranking (which has long premises).

We also repeat the experiment of evaluating the models on examples with varying premise lengths by using a combination of all the CFCS dev datasets described in Section 6. Figure 5 shows the considerably steeper decline in the performance of the MultiNLI model as compared to the RACE$_{Hybrid}$ model as the premise length increases beyond 200, similar to the trend observed on the task of MCRC (Figure 4).

The results of evaluations on both the downstream tasks provide sufficient evidence supporting our long premise conjecture. Moreover, since the
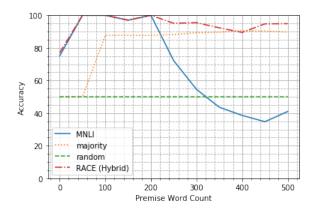

Figure 5: The graphs show the performance of models trained on RACE$_{Hyrbid}$ and MultiNLI at different premise lengths on a combined evaluation set of all the CFCS datasets mentioned in Section 6. . The accuracy at length x denotes the accuracy of the models on the examples with premise length in $[x, x + 50)$ words.

models trained on the converted data outperform all the reported results on the task of CFCS, they can act as a rich resource for the community for this task.

8 Conclusion

The difficulty of transferring entailment knowledge to downstream NLP tasks can be largely attributed to the difference in data distributions, specifically the premise lengths. Models trained on single-sentence NLI datasets are incapable of performing the multi-sentence inference required for the downstream tasks.

We experiment with three conversion strategies – rule-based, neural, and hybrid – to recast existing MCRC datasets into NLI datasets. We discuss the trade-off between structure and grammatical coherence in the context of the conversion, and perform experiments to identify the hybrid conversion strategy as the best. Recasting MCRC datasets into NLI using this strategy can result in broadly useful NLI datasets. Models trained on these multi-sentence NLI datasets perform better than models trained on existing single-sentence NLI datasets, in the context of the long-premise downstream tasks of MCRC and CFCS. These datasets can be a useful resource for creating truly generalizable NLI models.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large anno-

tated 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*.

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*.

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.

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 Reproducibility Checklist

A.1 Details of the datasets used

Table 8 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.

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 8: Number of examples in each of the datasets.

Table 9 shows the proportion (absolute numbers) of neural, rule-based and Q+A examples in the final hybrid datasets.

A.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 10 on google-sentence

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

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

completion dataset (Filippova and Altun, 2013)†

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

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 10: Hyperparameters and runtime metrics for training the neural conversion model

A.2.1 Experiments

- The hyperparams for the models used throughout the Section 7 are shown in Table 11. These were obtained using minimal manual tuning.

- The threshold for CFCS as classification experiments (Section 7.2.2 (1)) we calculated by tuning for best balanced accuaracy https://scikit-learn.org/stable/modules/generated/sklearn.metrics.balanced_accuracy_score.html.

B Conversion examples

Tables 12, 13 and 14 show examples of rule-based and neural conversions on RACE, MultiRC and DREAM respectively.

---
†https://github.com/google-research-datasets/sentence-compression
‡https://worksheets.codalab.org/worksheets/0xd4ebc52cebb84130a07cbfe81597aaf0/

	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 11: 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 7).

	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 12: 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 13: 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 14: Examples of Rule-based and Neural Conversions on DREAM