Reading Comprehension as Natural Language Inference: A Semantic Analysis

Anshuman Mishra [*1], Dhruvesh Patel[*1], Aparna Vijayakumar[*1],
Xiang Lorraine 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 the recent past, Natural language Inference (NLI) has gained significant attention, particularly given its promise for downstream NLP tasks. However, its true impact is limited and has not been well studied. Therefore, in this paper, we explore the utility of NLI for one of the most prominent downstream tasks, viz. Question Answering (QA). We transform one of the largest available MRC dataset (RACE) to an NLI form, and compare the performances of a state-of-the-art model (RoBERTa) on both these forms. We propose new characterizations of questions, and evaluate the performance of QA and NLI models on these categories. We highlight clear categories for which the model is able to perform better when the data is presented in a coherent entailment form, and a structured question-answer concatenation form, respectively.

1 Introduction

Given two sentences, a premise and a hypothesis, the task of Natural Language Inference (NLI) is to determine whether the premise entails the hypothesis or not. [†] The concept of semantic entailment is central to natural language understanding (Van Benthem et al., 2008; MacCartney and Manning, 2009) and therefore, NLI models have been used to help with various downstream tasks like reading comprehension (Trivedi et al., 2019), summarization (Falke et al., 2019; Kryściński et al., 2019), and dialog systems (Welleck et al., 2019). However, the performance of an NLI system on these downstream tasks has not been studied with respect to semantic or reasoning categories.

In this work, we use NLI to perform the task of multiple choice reading comprehension (MRC, or RC). We analyse the performance of an NLI model on this task through the lens of semantics by identifying the reasoning categories (type of questions) where it is beneficial to use an NLI model.

Drawing inspiration from the prior work in the area (Clark et al., 2018; Demszky et al., 2018; Trivedi et al., 2019), we use rule-based conversion to create an NLI version of the largest available RC dataset - RACE (Lai et al., 2017). We train a RoBERTa based RC model on the original dataset, and a similar RoBERTa based NLI model on the NLI version of the dataset. We evaluate and analyse the performance of both these models by characterizing the question types that are better suited for an NLI model and a QA model.

2 Related Work

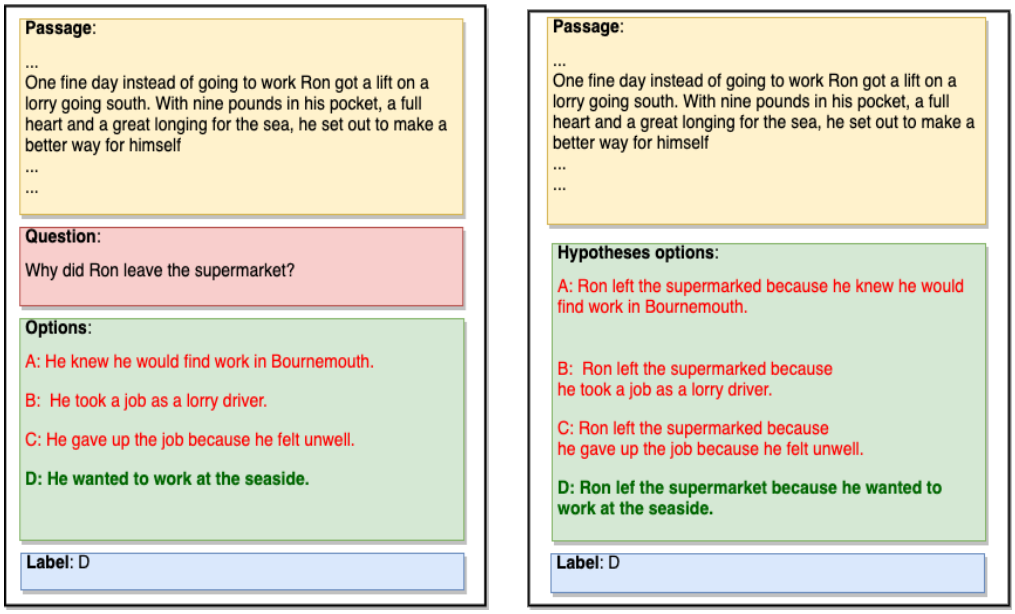
Reading comprehension (RC) is one of many potential downstream tasks that can benefit from NLI (MacCartney and Manning, 2009). It is easy to see that RC naturally reduces to a two-class NLI problem; specifically, it can be cast as the task of identifying if a given piece of text entails the statement formed by converting a question and a potential answer to an assertive statement (hypothesis).

Given the intuitive conversion between RC and NLI, Demszky et al. (2018) designed both a rule-based conversion system as well as a trained neural model to convert question-answering datasets such as SQuAD (Rajpurkar et al., 2016) and RACE (Lai et al., 2017) to an NLI form. However, it is unclear what these converted NLI datasets offer compared to the original question-answering datasets w.r.t semantics. We show that converting a RC task to an NLI task helps in answering certain types of

[*]Equal contribution.

[†]The "not entailment" can further be subdivided into "neutral" and "contradiction". However, we only use the two-class version of the problem in this work.

*Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 12–19
Barcelona, Spain (Online), December 12–13, 2020



(a) QA sample (b) Converted NLI sample

Figure 1: A RC sample with multiple answer choices converted to an NLI sample.

questions. This establishes the usefulness of the converted datasets.

Jin et al. (2019) show that despite the different form of NLI and QA tasks, performing coarse pretraining of models on NLI datasets like SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) not only improves the performance of these models on downstream reading comprehension tasks, but also helps with faster convergence. We show that – for certain types of questions in reading comprehension datasets – simply transforming the task to NLI can show improvement in performance, even without pretraining on any NLI dataset.

Trivedi et al. (2019) introduced a learnt weight-and-combine architecture to effectively re-purpose pretrained entailment models (trained on SNLI and MultiNLI) to solve the task of multi-hop reading comprehension. They show that for certain datasets, this strategy can produce good results. However, their study focuses mainly on improving model performance using a pre-trained NLI model, and lacks an analysis of the reasoning differences arising due to the different form of the NLI and QA tasks. We focus our analysis on this aspect.

3 NLI for Reading Comprehension

This section describes our experimental setup for comparing a QA based approach and an NLI based approach for the task of reading comprehension. We first obtain a parallel NLI and QA dataset by converting existing RC dataset into an NLI dataset. We then train two models, one on each form of the data, and analyse their performance.

3.1 Converting RC to NLI

We use the RACE dataset (Lai et al., 2017) for our experiments. It is a large-scale reading comprehension dataset comprising of questions collected from the English exams for junior Chinese students. We pick RACE because it is in a general domain and large enough to perform conclusive analysis. Each question in the dataset contains four answer options, out of which only one is correct. However, about 44% of the RACE dataset consists of cloze style (fill-in-the-blank) questions which are already in NLI form. Hence, in order to have a fair comparison, we only use the subset of RACE dataset which does not contain cloze style questions. This subset consists of 48890 train, 2496 validation and 2571 test examples.

We convert a RC example into an NLI example by reusing the passage as premise and paraphrasing the question along with each answer option as individual hypotheses as shown in Figure 1. Specifically, we generate the dependency parse of both the question and the answer option by using Stanford NLP package (Qi et al., 2018), then we follow the conversion rules proposed in Demszky et al. (2018) to generate a hypothesis sentence*. We make a few

*Appendix C presents example conversions generated us-

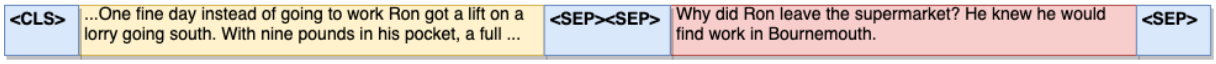


Figure 2: For the QA model, the input corresponding to each option contains question+option concatenation as the second sequence.

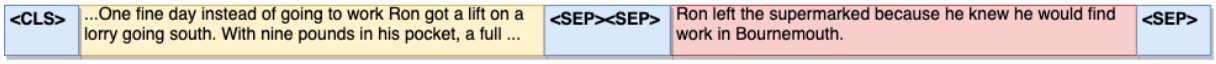


Figure 3: For the NLI model, the input corresponding to each option contains *hypothesis* generated using the question and option as the second sequence.

Dataset	Dataset Format	
	QA	NLI
RACE	**85.78**	-
RACE-subset	79.84	**82.09**

Table 1: Accuracy on the test set obtained by using different formats of the data.

additions to these rules to handle a some peculiar question categories in the RACE dataset. The most prominent of the added rules is the one for questions containing *"which of the following are (not) true"*. Such questions are very frequent (about 6% of all questions) and are not handled correctly by the rules in Demszky et al. (2018).

3.2 Model

In order to perform apple-to-apple comparison we use the same model architecture for both QA and NLI. Specifically, we use the state-of-the-art reading comprehension model – consisting of a RoBERTa model (pretrained on the masked language modeling objective) as the encoder and a two layer feed-forward network on its [CLS] token as the classification head – as described in Liu et al. (2019). The input sentence is the combination of the passage and its hypothesis. The hypothesis are created using the rule-based conversion method mentioned in Section 3.1 (NLI setup) or by concatenation of question and answer option (QA setup). The input to the QA setup and the NLI setup, corresponding to the first option in Figure 1, is shown in Figure 2 and 3.

4 Analysis

Table 1 shows the accuracy achieved by the RoBERTa model on the RACE dataset and its subset when presented in different forms. As we can see, the NLI model performs much better than the QA model on RACE subset. We think that the reason for this is the more natural form of the hypothesis statement used by the NLI model (Figure 3) compared to the Q+A concatenation form used by the QA model (Figure 2). Moreover, while the RACE-subset consists of only those questions which have question-words such as {*who, what, when...*}, about 95% of the rest of the dataset, i.e. RACE \ RACE-subset, consists of only fill-in-the-blank (FITB) type questions. These FITB questions are largely with the blank at the end of the question and a naive question-answer concatenation is very similar to a NLI form hypothesis. We believe that helps the QA model to perform better on the full RACE dataset compared to the RACE-subset, where NLI model is able to outperform the QA model showing the clear benefits of coherent conversion on complex question formulations (such as *W* word questions).

In order to analyse the performance difference from a semantic perspective, we characterize question into 7 semantic categories by identifying the kind of reasoning required to answer the question. Table 2 succinctly describes the reasoning categories.

4.1 Categories based on manual analysis

In order to perform manual analysis, we construct a *delta subset* consisting the 328 dev set examples on which the predictions of the QA and NLI models differ. We further divide the *delta subset* into *gain* and *loss* subsets. The *gain subset* consists of questions which the NLI model gets right, but the QA model gets wrong, and the *loss subset* is its complement in the *delta*.

We manually annotate these 328 (192 in *gain* and 136 in *loss*) examples into one of the 7 categories. However, about half the examples in the *delta subset* were not properly converted by the rules leading to unnatural or incoherent hypothesis

ing these rules.

Category	Description	Example
Linguistic Matching	Matching words between the question and a sentence in the passage	**Passage** : Food cooks quickly in parabolic cookers **Question**: If you want to cook food quickly, which kind of sun-cooker is your best choice?
Main Idea	Require topicality judgements	What's the best title for this passage?
Negation	Picking the incorrect statement	Which of the following statements is NOT true?
Dialogue	Can be inferred from a dialogue or direct speech in the passage	By saying "her pen dared travel where her eyes would not", the writer means
Math	Mathematically combining facts	How many functions of snow are discussed in the text?
Deductive	None of the above but can be answered precisely from the text	Which of the following statements is TRUE?
Inductive	None of the above and cannot be answered precisely from the text	How old is most likely the writer's father?

Table 2: Reasoning Categories (exclusive)

Type	Heuristics
Main Idea	Questions containing the words 'mainly', 'title', 'purpose' or 'topic'
Negation	Questions containing the 'not', 'except' or 'which of the following is wrong'
Dialogue	Passages containing more than 10 quotation marks (")
Math	Questions containing the words 'how many', 'how old' or 'how much'
Deductive	Questions containing the word 'true'

Table 3: Heuristically Determined Question Types in RACE-subset (non-exclusive).

sentences. Hence, for the purpose of illustration, we removed these examples leaving a total of 175 examples (109 in *gain* and 66 in *loss*) to do further analysis. Figures 4 and 5 show the distribution of labels over the Gain and Loss regions respectively. The distribution reflects that the QA models clearly outperforms the NLI model in negation questions whereas the NLI model outperforms the QA model in dialogue and deductive reasoning categories.

4.2 Categories based on heuristics

We also define another set of non-exclusive categories using heuristics, as described in Table 3. As shown in Table 4, the NLI model outperforms the QA model significantly in the dialogue, math and deductive reasoning categories. This overall trend further emphasizes the benefits of proper hypothesis generation as opposed to question and answer concatenation for the reading comprehension task.

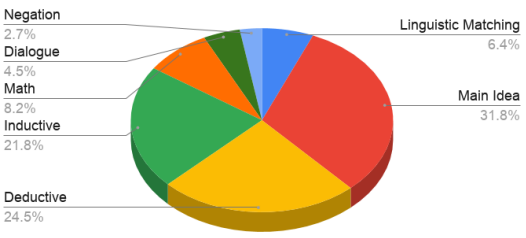


Figure 4: Reasoning categories of the gain region

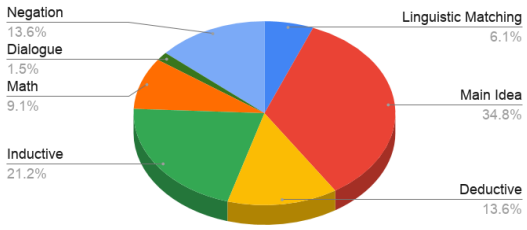


Figure 5: Reasoning categories of the loss region

Type	Fraction	QA	NLI
Main Idea	0.12	84.19	**84.83**
Negation	0.06	**80.86**	77.77
Dialogue	0.12	80.65	**83.60**
Math	0.03	45.00	**55.00**
Deductive	0.04	81.91	**88.29**

Table 4: Model performances on heuristically determined question types for RACE-Subset.

5 Conclusion

There is limited work providing a comprehensive analysis of how NLI can be used for QA. In our work, we show that NLI can be used for the task of reading comprehension simply by converting the data into NLI form. We convert a large RC dataset into NLI form and perform a comparative study of the performance of the RoBERTa model trained on QA and NLI settings. We propose a categorization of questions that allows for effective comparison of models trained on NLI and QA forms of data. Our analysis clearly shows that using the NLI-based approach is at par with a QA-based approach for most reasoning categories, and it is even better for some. Specifically, we find that questions involving deductive reasoning, dialogue interpretation and math are better handled by a model trained on the NLI form of data than the QA form. However, questions involving negation favor the QA form. Our work allows for careful selection of modeling strategy based on the type of data at hand.

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*.

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

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.

Di Jin, Shuyang Gao, Jiun-Yu Kao, Tagyoung Chung, and Dilek Hakkani-tur. 2019. Mmm: Multi-stage multi-task learning for multi-choice reading comprehension. *arXiv preprint arXiv:1910.00458*.

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

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

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*.

Bill MacCartney and Christopher D Manning. 2009. *Natural language inference*. Citeseer.

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*.

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.

Johan Van Benthem et al. 2008. *A brief history of natural logic*. LondonCollege Publications.

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.

A Model Architecture

Figures 6 and 7 show the model architecture for the QA and NLI models, respectively. As seen the architecture of the model is the same, the only difference is in the input form.

B Hyperparameter Settings

Hyperparam	**RACE-subset**	
	NLI-form	**QA-form**
learning rate	1e-5	1e-5
weight decay	0.01	0.01
warmup steps	1300	1300
batch size	16	16
max epochs	4	4

Table 5: Hyperparamter Setting

Table 5 lists the hyperparameter settings for both versions of the dataset.

C Conversion examples

Table 6 shows examples of NLI-form obtained applying rule-based conversion on QA examples from the RACE-subset.

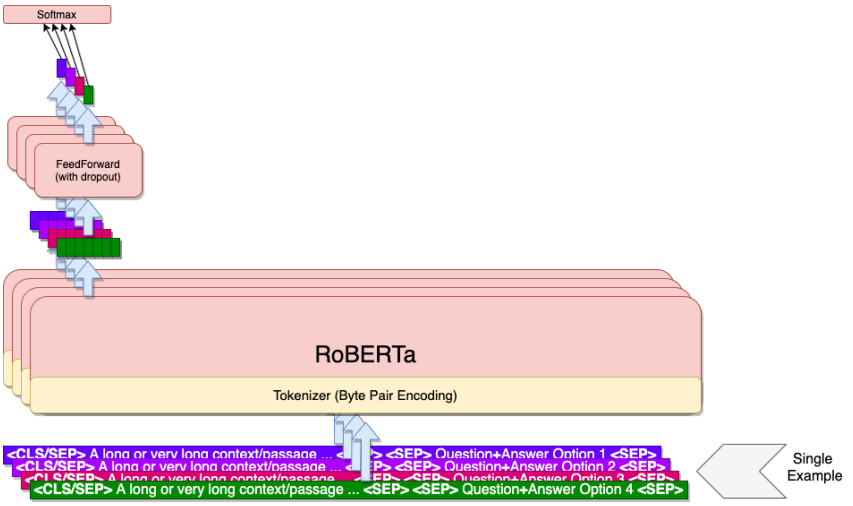


Figure 6: QA model

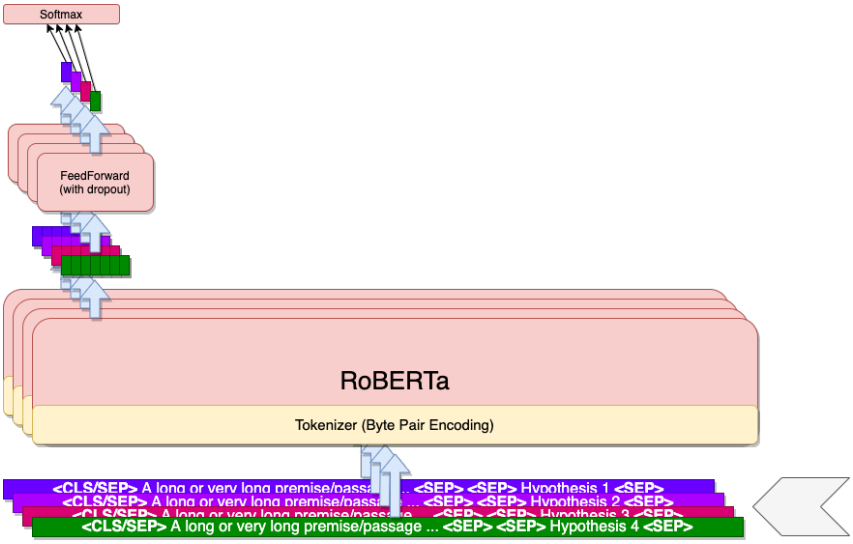


Figure 7: NLI model

QA example	NLI-form
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.
Q: What's the best title of the passage? **A:** Blame! Blame! Blame!	The best title of the passage's blame.
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.
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.

Table 6: Examples of Rule-based conversion applied to samples from the RACE-subset.