aNswER: A Named-Entity-Recognition based approach for MultiRC

Harshita Paila, Rohit Roongta, Soujanya R Bhat, Varun Chaudhary
{hpaila, rroongta, sranga16, vchaudh7}@asu.edu

Abstract

MultiRC(1) is a reading comprehension challenge in which questions can only be answered by taking into account information from multiple sentences. MultiRC task is one of the SuperGLUE tasks which provides a benchmark for evaluating general-purpose language understanding systems. *Jiant*(2), a modular software toolkit with configurable parameters, built on PyTorch, components from AllenNLP and the transformers package(9) implements and evaluates the baselines (3). In the first phase, we used the jiant toolkit to train, evaluate and analyze the results obtained for MultiRC. In support of an improved model performance, we locked down on an approach based on Named Entity Recognition (NER), meant for information extraction. In this phase, we have implemented **aNswER**,an NER-based mechanism for MultiRC task by processing the dataset from original dataset to a context(tokens and corresponding tags) based data with question-answers-paragraph concatenated, adjusting to our NER-pipeline. The performance was analysed and evaluated by taking different input variations, but judged on the same platform as jiant for comparison. Overall, our proposal of NER-based Question-Answering for MultiRC task has a better complexity(simpler), accuracy, confidence and performance. We compare and report supportive result as analysis for further proof of our findings.

1 Introduction

MultiRC dataset is mainly created to tackle the fundamental issue related to simple algorithms that can deal with the existing large datasets but fail at generalization. In this reading comprehension challenge, answering each of the questions requires reasoning over multiple sentences and each question can have multiple possible correct answers. The dataset is the first to study multi-sentence inference at scale, with a set of question types that require reasoning skills. SuperGLUE proposes an F1 evaluation metric over all answer-options ($F1_a$) to evaluate the performance of any model for this task. So, to keep the comparison platform common for our approaches, we decided upon F1 score as the discriminatory factor to determine the best model/approach.

Training Sample:


```
{
 "text":".<entire paragraph>",
  "questions":[
   {
    "question":"What do you apply to an object to make it move or stop?",
    "Sentences_used":[ 9, 5, 7 ],
    "answers":[
     {
      "text":"Strength",
      "isAnswer":false,
      "scores":{
      }
...
```


Figure 1: MultiRC data sample

2 Dataset Description

MultiRC is the dataset that is created to challenge the algorithms to attain deep understanding. This dataset consists of 800+ short paragraphs and approximately 6k multi-sentence questions extracted from 7 different domains such as news, fiction, historical text etc. Each question is associated with several choices for answer-options, out of which one or more correctly answer the question.The number of correct answers for each question is not predefined. These multiple variations and lack of pre-defined specifications such as variable options per question or variable correct choices per question provide a real challenge for any NLU model to learn patterns or representations in answering the questions of the dataset. On a subset of this dataset, human solvers are found to achieve an F1-score of 0.818 while BERT-large-cased achieved a score of 0.70. The distinguishing features of this dataset,restricting a model's capability to learn helpful information from any

pre-training task, includes:
1. Require multiple sentences to answer correctly
2. Multiple correct answers
3. Answers to be judged independently
4. Multiple source domains of information

3 Jiant(baseline): Implementation and Analysis

The model(2) implements a basic question-answering approach leveraging BERT as follows:

- Pre-processing: Loading and transformation of MultiRC dataset to adjust to the sentence-pair(in this case, context-question_answer pair) classification.
- Processing: The appropriate placement of *[SEP] and [CLS]* tokens, is followed by tokenization for the entire context.
- Training: The embeddings obtained from BERT(fine-tuned) are fed to the logistic regression classifier to predict the corresponding answer option as correct(label-1) or incorrect(label-0)
- Evaluation: Uses the ground truth labels to judge the model answer option predictions based on F1 metric score.

The evaluation metrics for the chosen model having the following characteristics(TP-true positives, FP-false positives, precision = $\frac{TP}{TP+FP}$, recall = $\frac{TP}{TP+FN}$):

- The model is observed to label a QA pair as True/1 if the answer consists of verbatim similar to the one present in the respective paragraph. This leads to a large number of predicted values to be True/1 as most of the answers provided in the task refer to a portion in the paragraph, which might not be relevant to the corresponding question.
- High Recall/Low Precision: The recall is expectantly high, as the model is more inclined to predicting the option as True/1(including a lot of FP), so there is a strong probability of high TP owing to this nature of model. The precision is adversely affected owing to large FP value.

PARAMETER	VALUE	PARAMETER	VALUE
max_seq_length	384	classifier	log_reg
batch_size	16	input_module	bert-base-cased
lr	0.00005	pretrain_tasks	sts-b,qnli
dropout	0.1	target_tasks	multirc
val_interval	1		

Figure 2: Jiant Parameters(chosen)

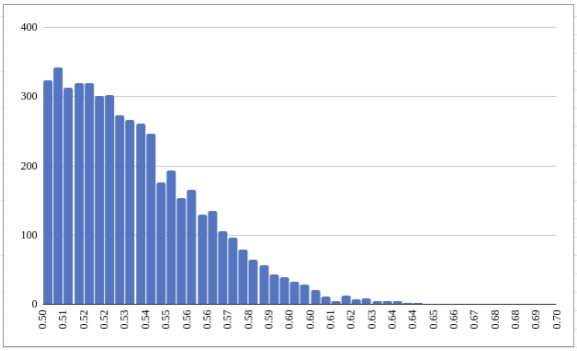

Figure 3: Jiant: Confidence histogram

For this phase, different models were run after choosing BERT as part of the baseline model, tweaking the hyper-parameters and leveraging pre-training tasks for fine-tuning to obtain the best possible results over the MultiRC dataset. There was minimal variation(approx. 2 units) in the $F1_a$ score. Since the model chosen by the reference paper(2) for providing baseline results(BERT-large-cased) was too resource-extensive for the available systems, BERT-base-cased was used throughout. The confidence probabilities paint a dark picture of the model, as highly under-confident with most logits ranging between 0.45-0.55, and even for relatively higher confidence area of 0.65 the model's predictions were not any better. In the figure, the red dots refer to an incorrect prediction while a green dot refers to a correct prediction. Also, the y-axis contains probabilities against the data points on the x-axis. This proved the approach to be an ineffective choice for the MultiRC task and we decided to experiment with a variety of different approaches elaborated further in the paper before choosing NER.

4 Alternative Approaches

We researched and analysed multiple approaches that were both a variation of the current baseline model methodology as well as a new direction to answering the dataset such as entailment and named-entity-recognition.

The first approach, we fine-tuned a BERT-base

Metric	Value	Metric	Value
TP	1913	Precision	0.4322
FP	2513	Recall	0.9219
TN	242	F1	0.5885
FN	162		

Figure 4: Jiant: Evaluation

QA model(on SQuAd) to derive answer spans from the paragraph. This approach failed due to the multi-hop property of the dataset, as the model was not able to perform multi-hop. This approach was inspired from a proposed relevant individual sentence-scoring mechanism to the question in *A Simple Yet Strong Pipeline for HotpotQA*(5).

The second approach, we implemented the pipeline recommended in an entailment-based approach(4). This approach considered the paragraph as premise, converted the question to hypotheses after incorporating each answer option individually and checking for entailment or contradiction. This provided good results with an F1 of approx. 0.63. We were able to add this approach as a separate task in the jiant toolkit for execution. While this approach was promising, the time-constraint and lack of personal novelty we did not explore further.

aNswER: The third approach is the focal point of this paper. We transformed the multi-hop multi-choice question answering approach to a named-entity-based task on the suggestion of our mentor. The steps taken are explained in detail below.

1. The first step was dataset transformation to an NER template. For this purpose, we wrote a script *parser.py* that performed the following operations:

a. Loaded the entire MultiRC dataset

b. Concatenated all parts of a data point in the order- *question* + *answer options* + *paragraph*.

c. Tokenized the entire context and allotted the following tags- P(for sentences in the paragraph). Q(for questions), C(for correct answers), W(for wrong answers) and I(as an inside tag).

d. Created a CSV file for the train and validation set with an ID column to uniquely identify every *(paragraph, question, answers)* combination.

2. The second step was the NER piepline that loaded, pre-processed, processed data and trained the *BertForTokenClassification* model to learn the tag representations.

3. The third step was executing the trained model on the validation set for evaluation (since the MultiRC test set labels are not publicly available, making it tough for result analysis).

4. The fourth step was evaluation, where only the C and W tags were analysed and compared with the predicted tags to compute the F1 score and model logits.

The **F1 score** and **confidence values** formed a common metric for model-model comparison among our implemented approaches.

This finalized NER approach provided a satisfactory F1 score over 63(the best we have observed working on the MultiRC dataset), after changing the problem statement entirely. Further analysis on the approach discovered a confident model with certain inaccuracies that were incorrectly penalised. Example- if a certain answer contains both C(correct) and W(incorrect) tags, the model is aiming to highlight only the relevant sub-parts. The high F1 score for the question tag also indicated that the model was able to learn the position of the question, which we believe was because of being at a similar position with respect to the end of the context and having a "?" tag at the end of itself.

Additionally, we looked at graph-based approaches for MultiRC- linked reasoning chain(6) and dynamically fused graphsfused-graphs. While graph seems to be a more intuitive option for a multi-hop task, we stuck to named-entity-recognition owing to its simplicity and novelty.

5 aNswER: Result Analysis

5.1 Data Preparation

For the dataset formulation, the following decisions were finalised:

- Order of concatenation: We finalised on the order as **question-answers-paragraph**. We experimented with paragraph-question-answers ordering and faced issues with respect to the trimming of last few tokens (which were answer tokens in the earlier case) due to maximum sequence length limitation of the BERT-base model. The average token length for a context after concatenation was approximately *350 tokens* which was well under our initial set length of 384. But, for the contexts larger than 384, the average was *450 tokens*. So, the answer options were trimmed and therefore the model did not predict any tags for those options. This issue was resolved by changing the ordering. A drawback is in-

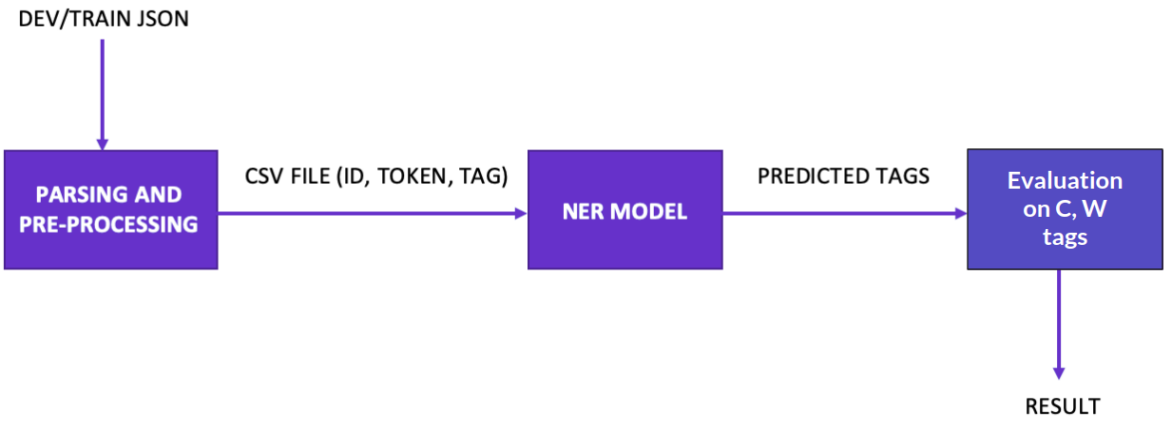

Figure 5: **aNswER**: Pipeline

Question word1	Question word2	Question word3	Question word4
Q	I	I	I
Correct ans word1	Correct ans word2	Correct ans word3	Correct ans word4
C	I	I	I
Wrong ans word1	Wrong ans word2	Wrong ans word3	Wrong ans word4
W	I	I	I
Paragraph word1	Paragraph word2	Paragraph word3	Paragraph word4
P	I	I	I

Figure 6: Tagging rules- Ground truth

PARAMETER	VALUE
Model	BertForTokenClassification
Batch Size	16
Max Input Length	400
Epochs	5
Optimizer	Adam
Learning rate	3 E-05
Loss Function	Cross Entropy

Figure 7: **aNswER**: Parameters

```
Accuracy score:
0.9865148169148586

              precision    recall  f1-score   support

           I     0.8816    0.9337    0.9069     17217
           Q     1.0000    1.0000    1.0000       953
           W     0.6785    0.8348    0.7486      2773
           C     0.6771    0.5123    0.5833      2075
           P     0.9097    0.9498    0.9293     11430

   micro avg     0.8554    0.9075    0.8807     34448
   macro avg     0.8655    0.9075    0.8847     34448
```

Figure 8: **aNswER**: Evaluation

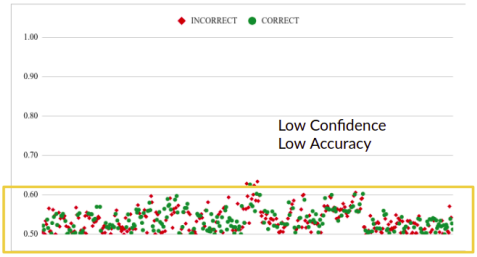

Figure 9: Jiant: Results(sampled)

deed the still existent trimming of the paragraph tokens now(instead of option tokens) but it is a model-based restriction and not an issue with the approach, with a relaxation that trimmed paragraph tokens might not always be important for the corresponding question. This can be resolved by using a BERT-large model.

- Tagging: We decided upon sentence-level tagging, with separate distinguishable tags each for- sentences in paragraph, question sentence, correct option sentence and incorrect option sentence. ”P”, ”Q”, ”C”, ”W” and ”I” tags were used to represent paragraph, question, correct option and wrong option and internal sentence words respectively.

5.2 Training

We fine-tune the *BertForTokenClassification* model on our training set consisting of 5100 data points(total answer options) which is pre-processed to a context associated with above mentioned tags. We run the training process for 5 epochs with a lr of 3e-5. We save our trained BERT model for predictions and later evaluation.

5.3 Evaluation

We evaluated the trained model on the MultiRC dev set with 953 data points(total answer options) and observed the metrics and model behavior reported via the figures 8,9,10 and 12. After plotting the confidence probability, it was observed a strong majority of **points denoting high probabil-**

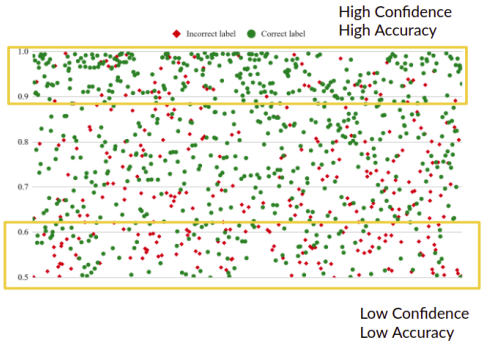

Figure 10: **aNswER** Results(sampled)

	Precision	Recall	F1
jiant	0.4322	0.9219	0.5885
MultiRC-NER	0.71	0.52	0.6

Figure 11: jiant v/s **aNswER**

ity had the correct labels whereas incorrect labels were observed primarily in the middle of the figure where the model was under confident. Additionally, there are only few points in boundary region for **aNswER** as opposed to the jiant confidence histogram where all points were enclosed within the boundary region of 0.45-0.55.

Additionally, analysing precision and recall in the various confidence intervals of 0.5-0.7, 0.7-0.85 and 0.85-1 gave us an insight into the model's pitfalls. **In the high confidence region(0.85+), the recall is lower** even in comparison to the previous intervals. We attribute this anomaly to the context lost while trimming down due to the sequence length limitation of our model(400 tokens, over which we dealt with resource limitations). The model confidently(and wrongly so) predicted a correct label as incorrect **because of lost tokens(and possibly relevant information) that it never knew existed**. We believe using an increased sequence length, and possibly a BERT-large model would provide a significant rise in recall in this confidence interval.

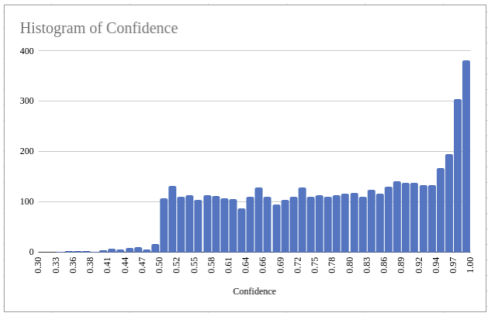

Figure 12: **aNswER**: Confidence Distribution

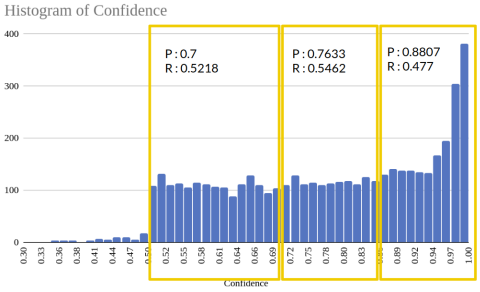

Figure 13: **aNswER**: Precision and Recall vs Confidence Distribution

6 Conclusion

After an extensive, but in no means exhaustive, analysis and implementation in terms of unique and traditional approaches to a multi-hop QA task, we believe the named-entity-recognition methodology is the most viable option, further proved by the highest F1 and confidence results. With a large number of variable factors in the dataset such as number of questions per passage, 2-4 sentences from the paragraph linked to answer a question, domains of information provided as context, a number of otherwise sensible approaches performed poorly, such as BERT-QA. We also realised that a model not only with improved F1 scores but also drastically enhanced confidence is a better option to work upon. The project also imparted the importance and scope of transfer learning. We ended up transforming the entire problem statement into completely new and resolving it with known tools. We also believe a more intelligent entity tagging approach and micro-analysis and rectification of the evaluation metric can pave the way for a significant model in the domain of multi-hop question answering.

7 Contributions

While the project was largely a team effort, tasks were divided along the way which everyone performed to the best of their abilities. We enjoyed working on state-of-the-art in the field of Natural Language Processing and the learning curve was an insightful experience with respect to multi-hop question answering. Each member's personal tasks and responsibilities are mentioned below, in addition to the joint effort on the entire project.

Harshita- Baseline model analysis with different hyper parameters, fine-tuning the parameters for choosing the best performed model; analysis of other approaches to QA including graph based ap-

proach and NER based question answering; NER-based QA training and evaluation notebooks; Result analysis based on F1-score.

Rohit- Baseline model analysis with different configuration variation, Fine-tuning the model parameters for understanding their effect on results, scrutinized numerous pre-processing mechanisms to improve the model accuracy, wrote NER pre-processing script to generate CSV file from JSON, CSV-file validation and result analysis of NER model.

Soujanya- Hyper-parameter tuning post baseline model exploration, baseline model analysis(codebase-level), implementation and analysis of Multee i.e. an entailment based approach to MultiRC, training and evaluation notebooks of NER-based QA approach and result analysis based on confidence probabilities.

Varun- Hyper-parameter tuning post baseline model exploration, baseline model analysis(confidence-level), analysis of BERT QA(fine-tuned on SQuAd) for multi-span answering, implementation and analysis of fine-tuned BERT models on tasks such as SQuAd, QNLI and STS-B; NER-based QA trainer and evaluation notebooks and result analysis based on F1-score.

References

[1] *Khashabi, Daniel Chaturvedi, Snigdha Roth, Michael Upadhyay, Shyam Roth, Dan. (2018). Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. 252-262. 10.18653/v1/N18-1023. (Paper)*

[2] *Pruksachatkun, Yada Yeres, Phil Liu, Haokun Phang, Jason Htut, Phu Wang, Alex Tenney, Ian Bowman, Samuel. (2020). jiant: A Software Toolkit for Research on General-Purpose Text Understanding Models.(Paper)*

[3] *Wang, Alex Pruksachatkun, Yada Nangia, Nikita Singh, Amanpreet Michael, Julian Hill, Felix Levy, Omer Bowman, Samuel. (2019). SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. (Paper)*

[4] *Trivedi, Harsh, Heeyoung Kwon, Tushar Khot, Ashish Sabharwal and Niranjan Balasubramanian. "Repurposing Entailment for Multi-Hop Question Answering Tasks." NAACL-HLT (2019). (Paper)*

[5] *Groeneveld, Dirk Khot, Tushar Mausam, Mausam. (2020). A Simple Yet Strong Pipeline for HotpotQA(2019). (Paper)*

[6] *Chen, Jifan, Shih-ting Lin, and Greg Durrett. "Multi-hop Question Answering via Reasoning Chains." arXiv preprint arXiv:1910.02610 (2019). (Paper)*

[7] *Xiao, Yunxuan, Yanru Qu, Lin Qiu, Hao Zhou, Lei Li, Weinan Zhang, and Yong Yu. "Dynamically fused graph network for multi-hop reasoning." arXiv preprint arXiv:1905.06933 (2019). (Paper)*

[8] *An Effective Multi-Stage Approach For Question Answering: Anonymous (Paper)*

[9] Huggingface Transformers github repository

[10] jiant tool-kit github repository

[11] CSE576 MultiRC project github repository

[12] An NER pipeline