Understanding stance of Editorial articles

Prajjwal Bhargava

October 2020

1 Methodology

Data The dataset consisted of 845 samples. Three samples were discarded because the tokenization output was greater than threshold value of 4096. The prompt is concatenated with each choice and padded to max length set to 4096. The training and validation set were randomly split across 842 samples with training and validation set consisting of 742 and 100 samples respectively. To take randomness into account, we provide results across 10 runs to get a better estimation. The prompt is concatenated with each options to generate 3 dimensional vector consisting of numericalized IDs, positional IDs and attention mask.

Model Since the prompt is huge (nearly all samples exceeds 1024 tokens), standard models such as BERT, RoBERTa cannot be directly utilized because their attention spans quadratically with sequence length leading to out of memory issues. We use Longformer, an extension of RoBERTa which uses global sliding window restriction attention to grow linearly with sequence length. Some attention heads rely on dilated attention to focus on global context while some heads focus on local context without dilated attention. We use the pre-trained weights provided with the paper. The model was pre-trained to process sequences of length spanning upto 4096 tokens. The model has a standard encoder similar to RoBERTa but with the mentioned attention mechanism along with a multiple choice head. The multiple choice heads receives scores for each option over which argmax is obtained to get the option with maximal confidence value.

2 Experiments

2.1 Random split

Figure 1 shows the results obtained on validation set for each run. Since the validation set is randomly determined, 10 runs give an approximate measure of how the model performs on similar distribution. These models are trained for 5 epochs. We observe that extended training does not necessarily leads to improved performance and early stopping can be used to address this issue. We use "body" as the main context from which options are be to predicted.

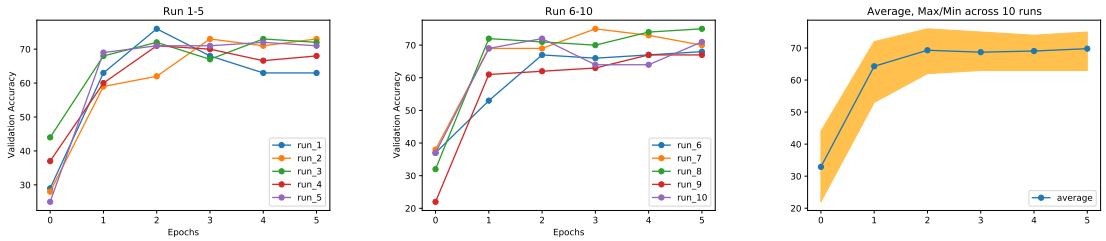


Figure 1: Accuracy of Longformer on Editorial articles dataset across 10 random splits (train-valid). For each run, the validation set is different

We randomly shuffle the options so that the network does not get biased to seeing the correct option at the same place during computation of loss. We perform truncation on sequences longer than 4096 (3 examples).

3 Cross validation (5 and 10 fold)

Figure 2 shows how longformer performs in 5 fold cross validation setting. In this case, the size of the validation set was kept to 169.

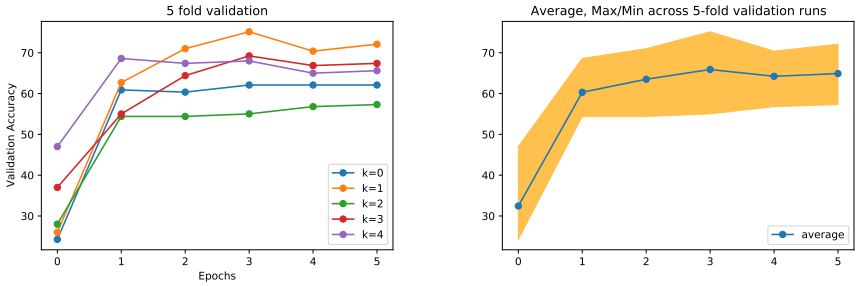


Figure 2: Accuracy of Longformer on Editorial articles dataset on all 5 folds including the average, min-max across 5 folds. For each k, the validation set is different

We also perform 10 fold cross validation with Longformer Figure 3. We observe that the model achieves high accuracy (>80%) on certain validation set. We also compute the average accuracy across all 10 folds ($\sim$ 70%).

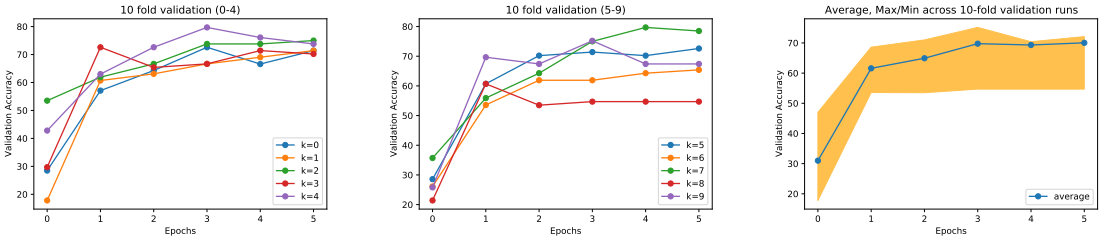


Figure 3: Accuracy of Longformer on Editorial articles dataset on all 10 folds including the average, min-max across 10 folds. For each k, the validation set is different

3.1 Evaluating with different contexts

In this section, we evaluate model's prediction when `headline` and `abstract` are used as a context in addition to `body` option. We are able to achieve substantially higher accuracy ($\sim$ 80%) by relying only on the abstract to predict the correct answers. Figure 4 shows the results obtained when only abstract is used with RoBERTa and Longformer. We observe that Longformer outperforms RoBERTa in case of 5 cross fold Cross validation.

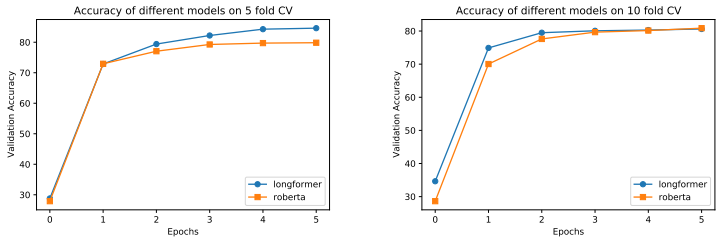


Figure 4: Accuracy of Longformer and RoBERTa on Editorials dataset averged out across 5 fold and 10 fold cross validation. We use **Abstract** as the context only.

When headline is used as the context to predict answers, we see that Longformer clearly outperforms RoBERTa.

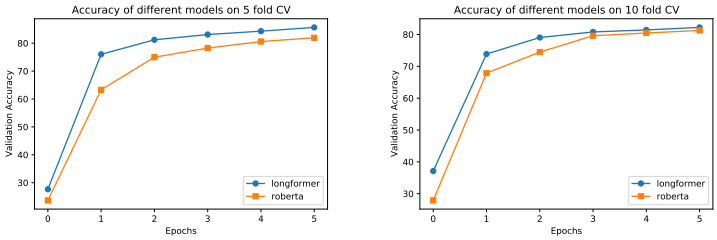
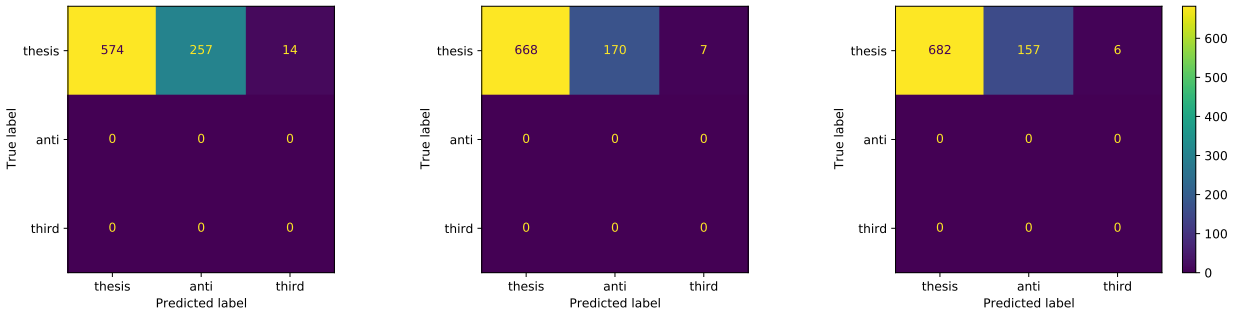


Figure 5: Accuracy of Longformer and RoBERTa on Editorials dataset averged out across 5 fold and 10 fold cross validation. We use **headline** as the context only.



(a) Body (b) Headline (c) Abstract

Figure 6: Confusion matrices for different contexts

3.2 Adding a neg-thesis option

In this section, we add a `neg-thesis` option which is contradictory statement of `thesis`. We evaluate if this can induce uncertainty in model's predictions.

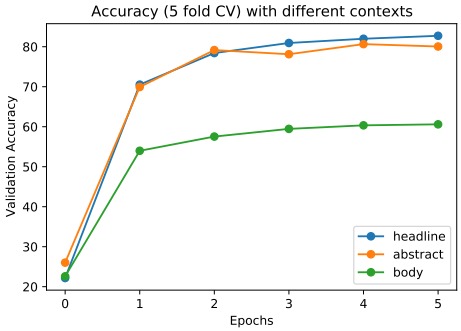
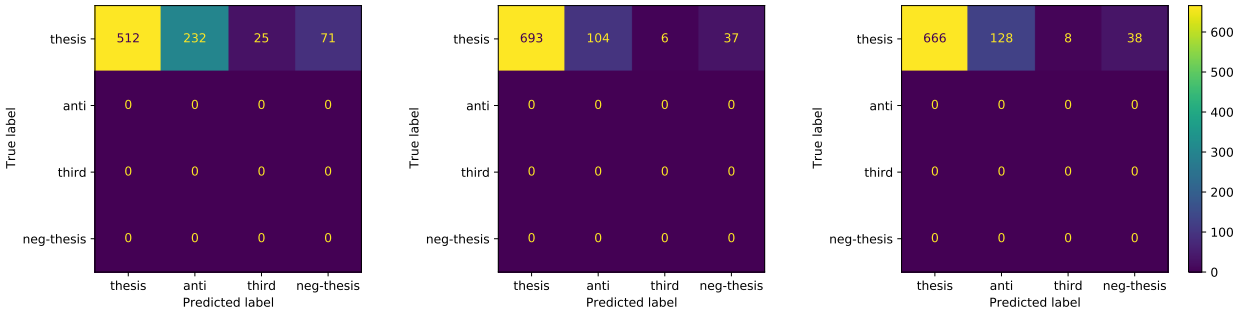


Figure 7: Accuracy of Longformer on Editorials dataset averged out across 5 fold.



(a) Body (b) Headline (c) Abstract

4 Replace options with a neural summarization LM

In this section, to induce more uncertainty and make provided options difficult, we replace `neg-thesis` (negation of thesis) and `third-option` (which was a random sentence derived from body), with the output from a neural summarization model. We use PEGASUS, state-of-the-art model for summarization. It was finetuned for paraphrasing task to make it output competitive sentences for short sentences. Without fineutning, we observe that these summarization models (BART, Prophetnet) output the similar sentences as the input possibly due to very short length of `third-option`.

4.1 neg-thesis

We find that synthetic `neg-thesis` reduces the uncertainty network has on other two class labels namely `third-option` and `neg-thesis` completely. The network seems to have learned the notion of negation in this context. Moreover it also has learned to distinguish between paraphrased sentences coming from body and what the body text actually implies.

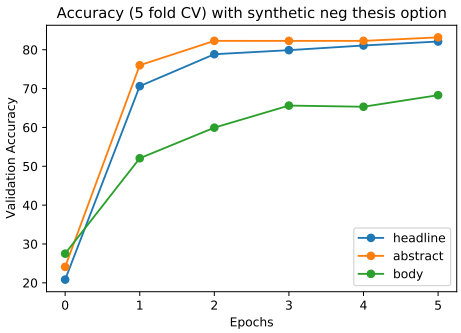
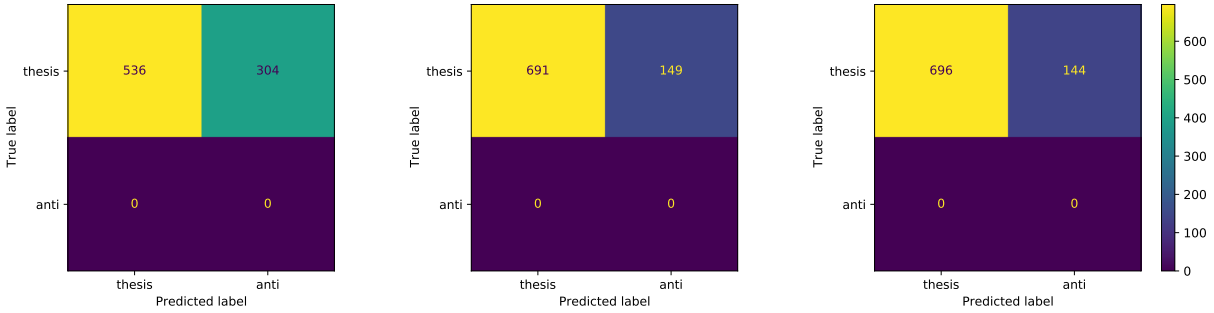


Figure 9: Accuracy of Longformer when `neg-thesis` is synthesized from PEGASUS averaged out across 5 fold.



(a) Body (b) Headline (c) Abstract

4.2 third-option

In this section, `third-option` is replaced by paraphrased sequences coming from PEGASUS.

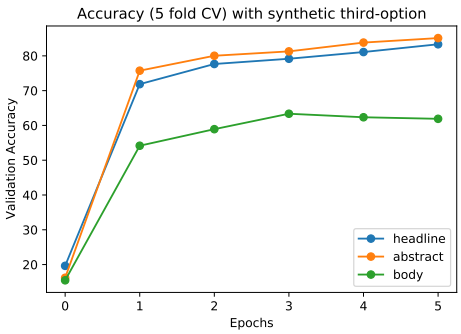
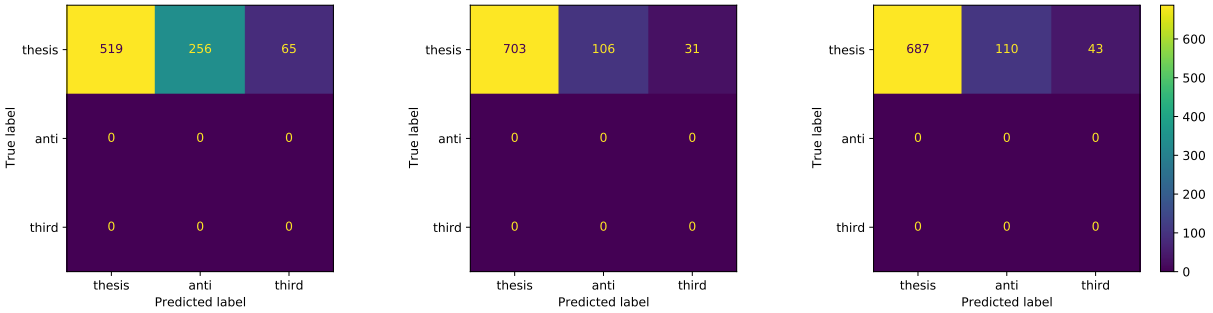


Figure 11: Accuracy of Longformer when `third-option` is synthesized from PEGASUS averaged out across 5 fold.



(a) Body (b) Headline (c) Abstract