Question Span Extraction from Chats of Instant Messaging Platforms

Stanford CS224N Custom Project

Abhishek Kumar
Stanford University
abhi2947@stanford.edu

Abstract

Instant messaging platform users often create random messages that are irrelevant to the community members to respond. We want to identify when a community member posts a relevant question so that we can alert the community experts (who can answer). To address this issue, we develop a novel data-set of 2925 span labels corresponding to the relevant questions in messages. We formulate two downstream natural language processing tasks separately - first, text classification to predict whether a message contains a relevant question or not, and second span extraction to output only relevant snippets of text. We adapt pre-trained models-BERT, SpanBERT, and RoBERTa for the classification task. We fine-tune, validate, and evaluate classifiers with 2400, 225, and 300 examples respectively. BERT(base) model achieves the best results with 85% accuracy, 96% precision, 80% recall, and 88% F1 score. We also introduce two customised metrics (Relevancy and Non-Relevancy Accuracy) that are more appropriate for our downstream application to evaluate classifiers. We introduce a new downstream task called 'extract relevant question span' to fine-tune T5 variants to extract text snippets corresponding to the relevant question in a given message. We also utilize pre-trained BART for span extraction task. Every span extraction model is fine-tuned and evaluated with 1200 and 300 examples respectively. Both types of models perform similarly with best BLEU-1 score of 0.83. BART-large and T5-base outperforms other models with 0.92 METEOR score.

1 Key Information to include

This project is mentored by Eric Scott Frankel with Vivek Khetan and Helix as external collaborators.

2 Introduction

Community-support platforms like Helix, give reward tokens to those community members who answer other members questions. Typically, the community members hang out on the instant messaging platforms like Discord and Telegram. We want to identify when a community member posts a relevant question so that we can alert the community experts (who can answer).

Users often create messages quite randomly. Not all the statements that sound like questions are questions. Some statements are the discussion around a question, which we want to ignore. While some messages contain relevant questions without punctuation marks, which we want to consider and extract. Table 3 includes a sample of annotation showing 4 different types of examples. For instance, T-4 example has a relevant question in the message without any question or punctuation mark, while T-3 example has a question but is not considered reasonable to be relevant and appropriate to pass on to experts to answer.

Through this project, we achieve two goals. First, we develop a novel data-set that includes the span labels corresponding to the relevant questions in the given chats sourced from the block-chain and

web 3.0 related channels of the instant messaging platforms. Table 3 includes a sample of annotation. The full data-set can be found here. Second, we utilize this data-set to explore different approaches to evaluate the performance of the existing state-of-the-art Transformer based models on the tasks of identifying and extracting the snippets corresponding to the relevant questions in posts of messaging apps.

3 Approach

3.1 Task Defining

This project identifies and extracts a relevant question in a text of messages through two natural language processing tasks to create an alert for the experts to answer. The term "relevant question" is defined as the span (part) of a text of messages or chats that is deemed reasonable to pass on for the pertinent experts to answer.

For a given text of messages, our first task is to classify the text whether it contains a relevant question or not, and the second task is to extract the span of text corresponding to relevant questions.We explore different approaches and existing state-of-the-art Transformer-based(Vaswani et al., 2017) models to solve these two tasks separately.

3.2 Methods

3.2.1 Text Classification

We adapt Transformer's(Vaswani et al., 2017) encoder-based models - BERT(Devlin et al., 2018), SpanBERT(Joshi et al., 2020), and RoBERTa(Liu et al., 2019) by adding a linear layer classifier to identify a relevant question in a text of message. The output layer of the customised models predict the probability of the given text has a relevant question. The model learns through Binary Cross Entropy objective function using AdamW optimizer. All pretrained parameters along with additional parameters are considered for training. Figure 1 shows the input-output pairs of the customised model.

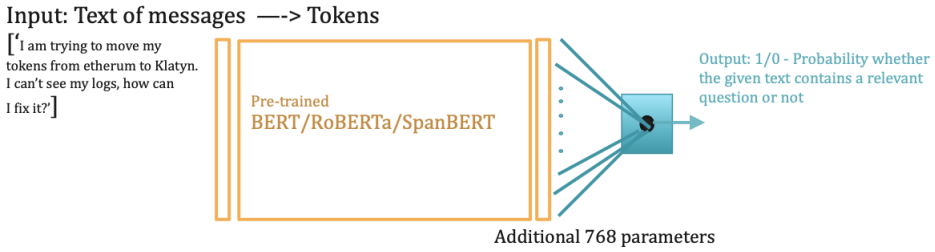

Figure 1: Input-Output pairs of the text classification models

3.2.2 Question Span Extraction

We utilize two kinds of pre-trained Transformer's encoder-decoder based models - T5 (Raffel et al., 2019) and BART(Lewis et al., 2019) to extract span of text corresponding to relevant question.

We define a new downstream task called "extract relevant question span" to solve span extraction task via fine-tuning customised T5Raffel et al. (2019) model. Table 1 includes a few examples of the input-output pairs that are used to develop T5 model. Figure 2 also highlights the working of the T5 for this task.

The pre-trained base and large version of BART is fine-tuned similar to the standard paraphrasing task to extract the relevant part of the input messages. Figure 2 shows the input-output pairs of the BART model.

Input	Output
'extract relevant question span: I am trying to move my tokens from etherum to klatyn. I can't see my logs, **how can I fix it?'**	how can I fix it?
'extract relevant question span: Maybe anyone know **how to import optimism snapshot into your node?**'	how to import optimism snapshot into your node?

Table 1: Examples of Input-Output pairs for T5 model.

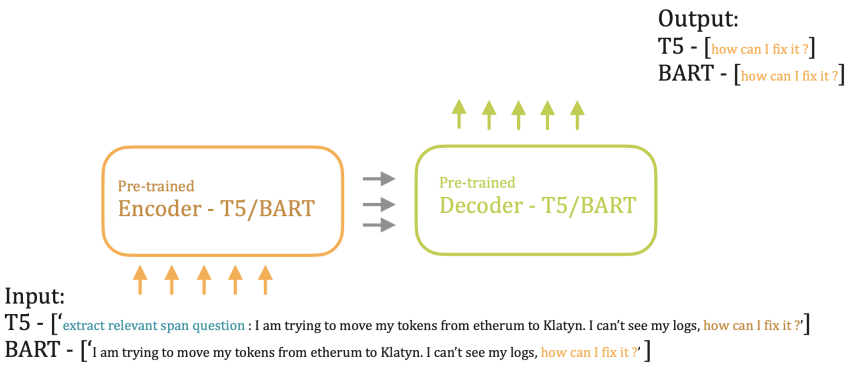

Figure 2: Input-Output pairs of the relevant question span extraction models

4 Experiments

4.1 Data

We collected and manually demarcated the contiguous segments of text corresponding to relevant questions of raw messages from Blockchain and Web 3.0 related channels of Discord and Telegram. Table 3 describes a sample of annotation that includes all 4 types of examples. Helix, a community-powered support company, provided all the resources for the data-set development.

Type	Description of Type	Train and Val	Test
T-1	Contain a span of relevant question	1500	194
T-2	Do not contain a span of relevant question	1125	106
T-3	Contain a question but not considered relevant	713	13
T-4	Contain a relevant question without any question mark	109	10
	Total	**2625**	**300**

Table 2: Summary of types of examples in the data-set for text classification task.

For text classification task, we fine-tuned and validated each model with 2400 and 225 examples respectively. We evaluated performance of the models using the metrics described in Section 5.2 with 300 examples.Table 2 describes the distribution of data-set. Of all test examples, 13 messages contain a question but are not considered relevant (to pass on to experts to answer), and 10 messages have no punctuation or question mark but have relevant questions within them that we want to identify and extract.

For span extraction task, we fine-tuned each model with 1200 examples that contain only relevant questions to extract them. We evaluated models with 300 examples.

4.2 Evaluation method

We evaluated text classification models with these metrics: - accuracy, precision, recall, and F1 scores. Accuracy is the fraction of predictions that are correct. Precision is a measure of how many of the

Type	Discord/Telegram Messages	Relevancy	Span Labels
T-1	I am trying to move my tokens from ethereum to klaytn. I can't see my logs, ***how can I fix it?***	1	how can I fix it?
T-2	Hi @louisli I've just given you a Guest Pass. You should be able to now	0	NO RELEVANT QUESTION
T-3	Hello there! Welcome to Klaytn Discord. How may I be of help today?	0	NO RELEVANT QUESTION
T-4	***Can someone guide me on how to get a guest pass*** pls	1	Can someone guide me on how to get a guest pass

Table 3: A sample of annotation including all 4 types of examples

positive predictions made are correct (true positives). Recall is a measure of how many of the positive cases the classifier correctly predicted, over all the positive cases in the data. F1-Score is a measure combining both precision and recall, which is the harmonic mean of the two.

For classification task, we define two new customised metrics for evaluation that are more appropriate for our project. First, we calculate the accuracy of model on those examples which include a question but are not relevant enough for our downstream application. We call it as "Non Relevancy Accuracy" and abbreviated as **"Accuracy-NR"**. Second, we estimate the accuracy on those messages which do not include any question/punctuation mark, but have relevant question span within the them. We call it as "Relevancy Accuracy", and abbreviated as **"Accuracy-R"**.

We calculated BLEU and METEOR scores to evaluate the performance of the span extraction task. BLEU, which highly correlates with human evaluation, calculates the proportion of how much of the generated N-grams in a candidate sentence got matched with the ground truth references. METEOR calculates the score by aligning generated captions to ground truth captions based on the accuracy and recall rate of the whole corpus. It also considers the form of words and language-specific resources that significantly improve the evaluation accuracy.

4.3 Experimental details

4.3.1 Text Classification

We imported pre-trained base and large models each of BERT, RoBERTa, and SpanBERT from huggingface transformers library and adapted as per the method described in Section 3.2.1 to classify a given text of messages whether it contains a relevant question. We fine-tuned and validated each model with 2625 examples, Binary Cross Entropy objective function, AdamW optimizer, 2e-5 learning rate, 32 input batch size, 8 epochs, using Pytorch-lightning in NVIDIA A100-SXM GPU. All the pre-trained along with additional parameters were considered for training.

4.3.2 Question Span Extraction

Similar to classification task, we imported pre-trained variants of T5 and BART from huggingface transformers library and adapted as per the method described in Section 3.2.2. We fine-tuned and validated each model with 1200 examples, Adam optimizer, 3e-5 learning rate, 0.1 dropout probability, 2 input batch size in NVIDIA A100-SXM GPU.

4.4 Results

Table 4 summarizes performances of 6 different fine-tuned models on our text classification task to identify whether the given input message contain a relevant question. BERT-base performs significantly better than all other models in terms of accuracy, recall, and F1 score. All models are absolutely precise with performance greater than 94%. Every model has recall notably lower than their precision and accuracy. Based on accuracy, BERT-base outperforms other models followed by RoBERTa-base. Base models of each variant perform better than large models considering accuracy,

recall, and F1 score. This could be due to larger models would require more data for fine-tuning to perform similar to base models with the same set of hyper-parameters in comparison to base models having significantly lesser parameters.

RoBERTa-large having least recall, performs notably better than other models in identifying those messages that contain questions but are not actual relevant questions and also those examples that include relevant questions without any punctuation/question mark. Performances of models based on Non-Relevancy and Relevancy Accuracy are discussed in detail in the Analysis section.

Text Classification Model	Accuracy	P	R	F1	Accuracy-NR	Accuracy-R
BERT - base	**0.85**	0.96	**0.80**	**0.88**	0.77	0.50
BERT - large	0.76	**0.98**	0.66	0.79	0.77	0.40
RoBERTa - base	0.81	0.96	0.74	0.83	0.84	0.60
RoBERTa - large	0.74	0.97	0.59	0.74	**0.92**	**0.70**
SpanBERT - base	0.75	**0.98**	0.64	0.78	0.77	0.40
SpanBERT - large	0.73	0.94	0.58	0.72	0.84	0.50

Table 4: Summary of Evaluation Scores of Text Classification Models

Table 5 depicts a summary of evaluation scores of 5 models on the task of extracting relevant question from a given message. All the models have similar performance in terms of BLEU score. Also, all N-gram BLEU scores are almost similar for each model, where $N \in [1, 4]$. Based on METEOR score, BART-large and T5-base have similar performance and better than other models.

Span Extraction Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
BART - base	0.83	0.83	0.82	0.82	0.85
BART - large	0.81	0.80	0.79	0.79	**0.92**
T5 - small	0.80	0.80	0.79	0.79	0.86
T5 - base	0.78	0.78	0.78	0.78	**0.92**
T5 - large	0.83	0.83	0.82	0.82	0.86

Table 5: Summary of Evaluation Scores of Span Extraction Models

5 Analysis

Table 4 depicts that all the classification models are very precise but have relatively low recall. Almost all the positives predicted by the models are correct but could not identify all the examples having relevant questions. Specifically, the models could not figure out well enough to those examples that have relevant questions without any punctuation/question mark. The example of such type (T-4) is shown in Table 3. This can be validated with low Relevancy Accuracy ("Accuracy-R") of each model shown in Table 4. RoBERTa-large performs the best with accuracy of 70% in identifying those messages that contain relevant questions without any punctuation or question marks. However, every model perform fairly well in classifying the messages that have any sort of questions but not relevant. Non-Relevancy Accuracy of each model can be observed in Table 4. It is also well noted from Table 2 that the test set contains only 8% of these (T-3 and T-4) examples. Results can be made more robust and reliable by evaluating the models with large test size with larger percentage of these two categories of examples.

BERT-base outperforms other classifiers in identifying the relevant question in chats in terms of half of the metrics considered for evaluation. This results are intuitively expected because BERT is pre-trained on the next sentence prediction task that is similar to our classification task. BERT generally performs better or similar in text classification task in comparison to other existing models. SpanBERT is not pre-trained on next sentence classification task as the original BERT. This could be a probable reason of the better performance of BERT than SpanBERT for our task.

All variants of BART and T5 performs fairly well in extracting the snippets of the text corresponding to the relevant question in the given message. As can be seen in Table 5, every model's performance is consistent based on all N-grams BLEU score. All models perform better based on METEOR score than BLEU. BART large and T5-base outperforms other models. BART and T5 models are

encoder-decoder models that achieve state-of-the-art results in the standard NLP summarising and text generation tasks. These models are also pre-trained on the similar tasks like ours with huge amount of data. These fine-tuned models deliver satisfactory results on our designed downstream task of extracting span of text corresponding to the relevant question in the chat.

On analysing the predicted output of the span extraction models on the test set, two major observations are noted. First, predictions of all the models are either very close or greater in span length than the human annotated label in more than 90% of the test examples. But the span prediction greater than the expected should not be an issue for our downstream task. Second, BART-base model looks to miss predict the initial question related words like - how, can, what, etc. in more than 50% of the test examples. Table 6 shows some examples of the predicted and actual labels. However, BART-large and all variants of T5 does not seem to have this issue. This could probably be solved by iterating the model for longer duration or by feeding with more training data, because BART base has only 35% of parameters than their large variant and hence less learning capacity based on the similar set of fine-tuning hyper-parameters and data credentials.

Human Annotated Label	**BART-base Prediction**
does anyone know if there is a way to get OP on the Goerli network?	anyone know if there is a way to get OP on the Goerli network?
can I withdraw OP tokens directly to Binance international for example?	I withdraw OP tokens directly to Binance international for example?
what alchemy provide public free api? not need to register?	alchemy provide public free api? not need to register?

Table 6: Comparison of BART-base prediction with human annotated label on test examples

6 Conclusion

This project attempts to address the issues of skipping redundant span of texts in the chats of the instant messaging platforms and identifying the relevant questions within them. The application of the task is to create an alert for the community members when a relevant question comes up in the Discord/Telegram channel so that they could contribute to help the members of the community or channel. We develop a novel data-set that contains 2925 examples of annotations including two kinds of labels - first, snippets of text corresponding to the relevant question in the messages collected from the block-chain and web 3.0 related Discord channels, and second, binary label to classify whether the messages contain a relevant question or not. We further plan to add more examples in the database covering from a wide variety of block-chain and web 3.0 related channels and make it publicly available for the utilization of the development of models on various downstream natural language processing tasks.

We formulate two tasks - first, text classification task to identify the relevant question in the messages, and second, relevant question span extraction task. We adapt pre-trained BERT, RoBERTa, and SpanBERT by adding a linear layer to predict the probability of the given text as having relevant question span or not, and then fine-tuned, validated, and evaluated with our customised data-set. BERT-base outperforms other models in terms of accuracy, recall, and F1 score. Although, all models struggle to identify the examples that have a relevant question without a punctuation or question mark, RoBERTa-large performs relatively better.

For span extraction task, we utilized pre-trained T5 and BART. We introduce a new task called 'extract relevant question span' and fine-tuned T5 model to predict the relevant snippets. We also fine-tuned BART-base and large like the standard paraphrasing task. Both models deliver satisfactory results for our downstream application. The best model (BART-large) has a METEOR score of 0.92 and a BLEU-1 score of 0.81.

Based on the development of the customised models with this amount of data to carry out our defined tasks, we realise that models learn sufficiently to deliver the tasks, but the evaluation results would be

concluded more concretely if we would use larger test set with inclusion of types (T-3 and T-4) of examples on which models tend to work less accurately than in general.

References

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

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.