ANLP Project: Final Report

Akshit Sharma
Rudransh Agarwal
Sai Asrith Devisetti

November 2025

Team-6: Phrase Phantoms
Title: Query-Guided Multi-Perspective Answer Summarization

1 Introduction

Text summarization is the technique of transforming long documents into shortened versions, while focusing on the sections that convey useful information, and without losing the overall meaning. Query-focused summarization is a sub-task within text summarization that aims to generate a summary of given text conditioned upon a user-query that is passed alongside the source document as input to the model.

Query-focused summarization is also relevant for Community Question Answering, where a person poses a question and can get multiple answers to sift through. Work in this field has a notion of 'best answer' and make use of this best answer as the gold summary of all other answers. However, best answer only presents one perspective and rarely captures the variety of perspective of other answers. Ideally, an answer summary should cover multiple perspectives found in answers. Answer summarization is a form of query-based, multi-document summarization, and creating answer summaries that reflect the underlying varying perspectives entails several sub-tasks:

- Selection of answer sentences relevant to the question (query sentence relevance)
- Grouping these sentences based on perspectives (clustering)
- Summarizing each perspective (cluster summarization)
- Fusing clusters into a coherent, overall summary (fusion)

2 Problem Statement

In this project, we focus on generating a fluent and concise answer summary that includes perspectives of all the answers for a given query on a community question-answering forum. To break down the problem, query-guided refers to using the question as a guide for summarization, while multi-perspective refers to using the perspective from each of the multiple available answers to produce a single summarized answer.

3 Data Collection and Pre-Processing

We are required to work with the **AnswerSumm** dataset, proposed in the paper titled **'AnswerSumm: A Manually-Curated Dataset and Pipeline for Answer Summarisation'** (Paper Link). It is a English-language dataset of questions and answers collected from a StackExchange data dump. The dataset was created to support the task of query-focused answer summarization with an emphasis on multi-perspective answers. The dataset consists of 4631 such question-answer threads annotated by professional linguists and includes over 8700 summaries. For each thread, the annotator writes two summaries. In First Summary, the annotator is asked to mark sentences that are included in the final summary and instructed to more closely use the words in these sentences rather than abstract. In Second Summary the annotator was asked to paraphrase and condense the cluster summaries but was not asked to reduce abstraction. The AnswerSumm dataset is readily available through

Question: I recently relocated to USA and have no Credit Score. Is Secure Credit Card is the only option for me to start building my credit score? Also please recommend which other credit cards are available for people like me to build credit score
Answer 1: If you have an AMEX from another country, you can get an AMEX in the US. American Express has a separate system that is not as strongly country-dependent as, say, VISA and MasterCard...
Answer 2: Secured credit cards are usually not very cost effective for building credit. Find a local credit union, of medium to large size. A credit union is like a bank, but operates under slightly different rules, and is non-profit...
Answer 3: If you have had an American Express card abroad, you can try and get a US Amex...
Answer 4: If the country you came from has an HSBC, you can ask HSBC to use your credit rating from that country to give you an HSBC Mastercard in the US...
Summary:
There are a range of options available to you, although your chance of success will depend on the bank that you apply with. However, if you have previously had a card with HSBC or American Express, the process may be simpler. Other options could include borowing from a credit union or asking a friend or family member to be an additional cardholder with you.

Figure 1: An example summary from the AnswerSumm dataset, illustrating the multiple viewpoints present manually-written summaries, and a subset of the 8 user answers to which the summary can be aligned.

the Datasets library from Hugging Face. We used Datasets library in our implementation

as well. After loading, we performed pre-processing steps like separating the train, test and validation sets, extracting the required entries (like question, answers, cluster summaries, first and second summary, etc.) for our use-case.

4 Baseline- End-to-End Approach (Mid-Submission)

BART (Bidirectional and Auto-Regressive Transformer): It is a Transformer-based model introduced by Facebook, designed for sequence-to-sequence tasks, such as text summarization, machine translation, and text generation. BART combines the best features of both bidirectional (like BERT) and auto-regressive (like GPT) models.

4.1 Experiments

For our baseline, we used BART for end-to-end summarization as it is trained on the summarization task. Since the input token limit for BART is 1024, any larger input was truncated. The output length was limited to a maximum of 256 tokens. We ran the following experiments:

- **AllAns+Pre-Trained BART**: We concatenated all the answers and used it as input to the model. The output summary was compared against the true abstractive summary, available for the samples, by calculating the ROUGE scores.
- **Query+AllAns+Pre-Trained BART**: Along with the concatenated answers, the query was also prepended to the input and passed into the model. The ROUGE scores were slightly higher in this case. We noticed that the output summaries contained words from the query which were also present in the true summaries, leading to a higher ROUGE overlap.
- **AllAns+Fine-Tuned BART**: We fine-tuned BART using the train set samples of the AnswerSumm dataset, without the query in the concatenated answers as input. We saw a significant improvement in ROUGE scores with fine-tuning, compared to the pre-trained model.
- **Query+AllAns+Fine-Tuned BART**: We fine-tuned using the train set samples of the AnswerSumm dataset with the inputs containing the query along with the concatenated answers. The aim was to make the model learn how to include multiple-perspectives given in the input answers, in the summary. We see a slightly higher ROUGE-2 and ROUGE-L score, most probably because of query tokens being included in output summary.

Experiment Type	ROUGE-1	ROUGE-2	ROUGE-L
AllAns+Pre-Trained BART	16.09	4.84	10.06
Query+AllAns+Pre-Trained BART	16.34	4.79	10.27
AllAns+Fine-Tuned BART	23.60	4.81	14.46
Query+AllAns+Fine-Tuned BART	18.66	4.89	14.59

Table 1: Average ROUGE scores (test set) for baseline experiments

5 Query Guided Summarization Approaches

We experimented with 3 different techniques, each focusing on the QFS (Query Focussed Summarization) task. The aim was to use only the sentences which are relevant to the query for generating the summary. All were two-step approaches, the first step being the selection of query-relevant sentences to create the input for the model and the second step of using this for generating the final summary using BART, fine tuned on AnswerSumm (train set) on summarization task.

5.1 Single Encoder Model: RelReg

RelReg (Relevance Regression) is a model trained to predict the ROUGE overlap between a source passage and the reference summary, using only the passage and query as input. A single-encoder model jointly encodes the delimiter separated query and passage, and the final layer of the model outputs the predicted relevance value. Here, ROUGE score is used as a proxy for relevance. The source passage is created by concatenating all the answers and the reference summary is the abstractive summary for the sample. We use BERT-base as the encoder model. For a given query+sentence input to RelReg, we use the model's predicted score to rank the sentences by relevance. We then experimented by picking top-k (k=5,10 and 15) highest scoring sentences as input to BART. We observe a significant improvement in the metrics for the test set, compared to the baseline.

5.2 Dual Encoder Model: RelRegTT

RelRegTT (Relevance Regression Twin Tower) is a more computationally-efficient version of RelReg that uses a dual-encoder architecture to predict ROUGE-based relevance scores. A sentence-encoder model is used to generate sentence embeddings for all the sentences in the answers. Then, the cosine similarity score between the query and sentence embeddings is used as a proxy for relevance of a sentence to the query. We rank the sentences by this score and experiment with top-k (k=5,10 and 15) highest scoring sentences as input to BART for generating the summary. We also used euclidean distance as our metric for scoring and ranking and found the results to be pretty poor, compared to cosine similarity (and even poorer than baseline of pre-trained BART, end-to-end summarisation).

multi-qa-mpnet-base-cos-v1: It is a pre-trained sentence transformer model available on Hugging Face and part of the sentence-transformers library. It is designed for semantic search and question-answering tasks. The model is based on the MPNet architecture and fine-tuned specifically for use cases requiring multi-modal or multi-source question answering.

5.3 Relevance Classification Model: RelClass

Similar to the relevance regression (RelReg) model, we came up with the RelClass (Relevance Classification) model, which classifies a sentence as relevant or irrelevant for a query. From the AnswerSumm dataset, we have the cluster ids available for each sentence. A cluster id of -1 indicates that the sentence was not a part of any cluster or perspective and can thus be used as negative sample (irrelevant sentence) for the classification model. A sentence with any other cluster id is used as a positive sample. We use BERT-base to encode the query+sentence and the output from the final layer is mapped to 0 (for irrelevant sentence)

and 1 (for relevant sentence). Then, while testing, we use only the sentences which the model predicts as relevant and as input to BART for generating the final summary.

6 Query Guided+Multi-Perspective Summ. Approach

Even though the dataset contains cluster ids for sentences, they are valid only for the sentences that have been used for the summary, instead of assigning clusters to sentences after performing overall clustering. So, due to lack of training data, we use the unsupervised clustering approach of DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to form sentence clusters, after encoding each sentence using a sentence-encoder, and using the clusters for generating a multi-perpective summary for the query. The idea is that after clustering, each cluster would represent a perspective and the outliers will be the sentences that are irrelevant for the query and thus not required for summary generation. Within the clusters, we rank the sentences by using the RelRegTT approach and use only top-k (k=3,5) sentences from each cluster to create the input for BART (fine-tuned), which then generates the summary. The multi-persepective nature of this approach comes from clustering (and amplified by the fine-tuned BART) and query-guided nature comes from the score-and-rank part (RelRegTT).

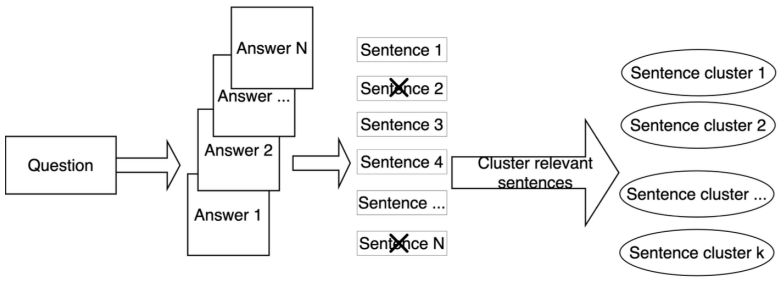

Figure 2: An illustration of our clustering approach

7 Observations and Results

- Table-2 contains the metrics obtained by running the models described above on the test-set of AnswerSumm dataset.
- BART is not trained for the task of query focused summarization. Hence, when we use BART for summarization, with and without query, the metrics are not very different. Whereas, when we fine-tune BART with the train dataset, without the task of query focused summarization, but just for fine tuning on the dataset, we see some improvement.
- Among the extractive-abstractive algorithms, the best performing model is RelRegTT which ranks sentences with similarity between query and sentences, from different models.

- Our best performing approach is the Unsupervised Clustering based approach followed by a ranking of sentences. Reason- while the query focused summarization methods work well on bigger documents, where there could be a lot of irrelevant information, our dataset consists merely of answers from a question. Thus, almost all sentences consist of relevant information, unless they are a part of a bigger answer with trivial information or they belong to a irrelevant answer. These sentences can be reasoned as being classified as outliers. Moreover, since our task is to perform multi perspective summarization, each cluster can be reasoned as representing one perspective, and thus, we include all the perspectives by including sentences from each cluster.

Technique	ROUGE-1	ROUGE-2	ROUGE-L
RelReg top-5	21.64	5.60	16.28
RelReg top-10	22.81	6.19	17.03
RelReg top-15	23.44	6.66	17.62
RelRegTT top-5 (cosine)	23.14	6.54	17.37
RelRegTT top-10 (cosine)	23.93	7.00	18.02
RelRegTT top-15 (cosine)	24.21	6.90	17.94
RelRegTT top-5 (euclidean)	10.62	1.16	8.47
RelRegTT top-10 (euclidean)	13.91	2.10	10.72
RelRegTT top-15 (euclidean)	16.18	3.01	12.27
RelClass top-5	21.47	5.85	16.37
RelClass top-10	22.29	6.19	16.80
RelClass top-15	22.39	6.32	17.02
RelRegTT+Clust. top-3	24.21	6.91	17.97
RelRegTT+Clust. top-5	25.35	7.35	18.61

Table 2: Average ROUGE scores (test set) for different approaches

8 References

- Multi-Perspective Answer Summarisation Paper: AnswerSumm: A Manually-Curated Dataset and Pipeline for Answer Summarisation
- Query Guided Summarisation Paper: Exploring Neural Models for Query-Focused Summarization
- multi-qa-mpnet-base-cos-v1 (Model card from Hugging Face)
- bert-base-uncased (Model card from Hugging Face)
- Facebook's BART-large (Model card from Hugging Face)
- AnswerSumm (Dataset card from Hugging Face)