Submodular Context Selection for Long-Document Question-Answering

CS769 Course Project Presentation (2023)

Adithya Bhaskar (190050005, adithyabcse@gmail.com)

Harshit Varma (190100055, varmaharshit2@gmail.com)

Guide: Prof. Ganesh Ramakrishnan

Long-Document Question-Answering

Every man's mind is a universe with …

•••

"I will meet you there in an **hour**," **she** said.

•••

Ever since he had first **set foot into his mind**, some **ten hours ago**, …

•••

It was a remarkably detailed materialization, and his quarry's footprints stood out clearly in the duplicated sand. **Sabrina York** did not even know the rudiments of the art of throwing off a mind-tracker. It would have done her but little good if she had, for **twelve years** as a psycheye had taught Blake all the tricks.

•••

He picked up **Sabrina**'s trail in the backyard and followed it down to the Martian waterway and thence along the bank to where the waterway ended and a campus began.

How much time has passed between Blake's night with Eldoria and his search for Sabrina York in his mind-world?

(A) 7 Years (B) 10 Hours (C) 12 Years (D) 1 Hour

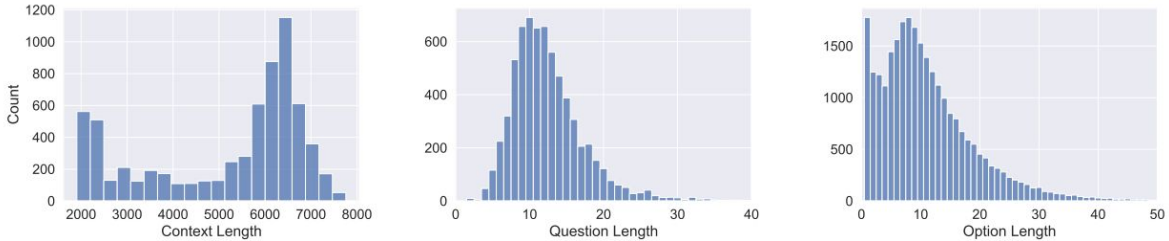
Related Work

- **Efficient Language Models**
  - Longformer (Beltagy et. al.)
  - Reformer (Kitaev et. al.)
  - Treeformer (Madaan et. al.)
  - Flash Attention (Dao et. al.)
- **Selector-Predictor Models**
  - REINFORCE-based sentence selector for event relation extraction (Man et al.)
- **Submodular Subset Selection in NLP**
  - Document summarization (Lin et al.)
  - Conditional data summarization (Kumari and Bilmes)
  - Diverse paraphrasing and data augmentation (Kumar et. al.)

Related Work

To the best of our knowledge, no prior work has studied submodular optimization in the context of Question Answering

QuALITY (Pang et. al.)



(Figure from Pang et. al.)

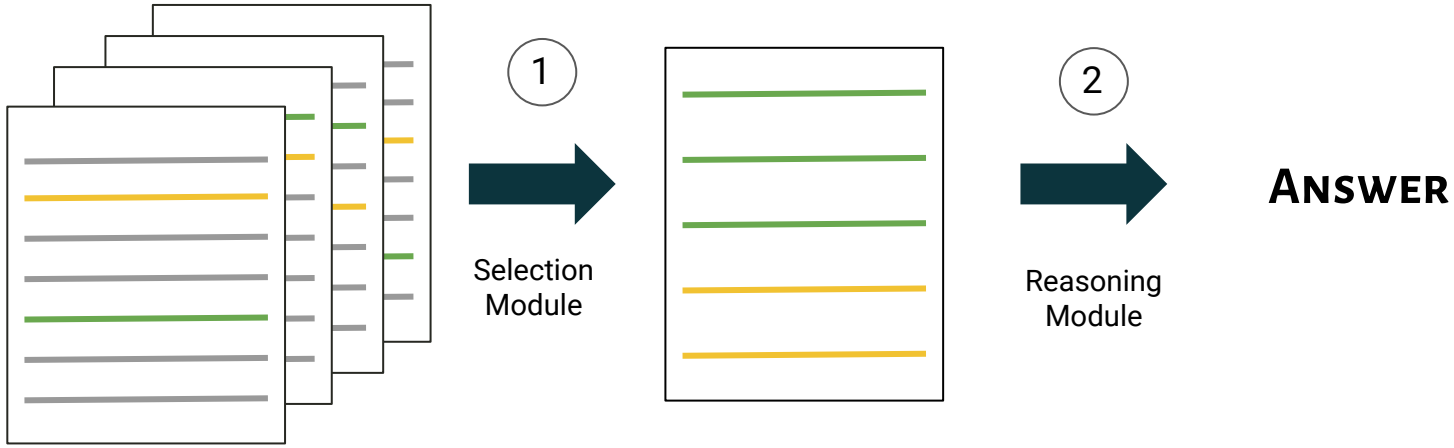
Figure 2: Article length, question length, and option length in QuALITY. The average length of an article, question, and option is 5,159 tokens, 12.5 tokens, and 11.2 tokens, respectively. The maximum length of an article, question, and option is 7,759 tokens, 103 tokens, and 75 tokens, respectively. The histograms are truncated to only keep the visible mass.

Split	No. of Articles	Avg. Article Length (Characters)	No. of Questions
train	300	24348.86	2523
dev	230	24306.59	2086

Table 1: Characterisitcs of the QuALITY dataset. Contexts often exceed 20,000 characters in length.

- ~6000 tokens (5000 words) takes approximately ~30 mins to read
- The dev split is further divided into the dev-dev (“dev”) and dev-test (“test”) splits in a 30:70 ratio (with 625 and 1461 questions each)

Our Approach



Selection Module

- Aim
  - Select sentences similar to the query
  - Cast a wide net in case we happen to miss
- Submodular Functions
  - Mutual Information Based Objectives allow fine-grained trade-off
  - Objectives are submodular irrespective of similarity function
  - Efficient, provably near-optimal performance of simple (greedy) algorithms
- Embeddings
  - Each span of up to 5 sentences
  - Facebook DPR (**D**ense **P**assage **R**etrieval)
  - Embeddings based on Sentence Transformers, TF-IDF, BM25 also supported for all objectives
  - Cosine similarity used as the similarity metric

Selection Module - Objectives

- Graph Cut Mutual Information
  - Functionally equivalent to greedy selection of most similar sentences
  - *Highly* prefers query relevance to diversity
- Facility Location Mutual Information
  - *Highly* prefers diversity to query relevance
- Log Determinant Mutual Information
  - Prefers query relevance to diversity
- Concave-over-Modular
  - Square-root as the Concave Function
  - Somewhat balanced between query relevance and diversity
- Baselines
  - No context, Random Selection, TF-IDF greedy, and BM-25 greedy

Selection Module - Objectives

- Syntactic Fidelity (Kumar et. al.)
  - Based on the *n*-gram overlap between the query and context sentences
  - Not implemented in `submodlib`

$$f(X; q) = \sqrt{\sum_{x \in X} \sum_{n=1}^{n} \beta^n \left| x_{\text{n-gram}} \cap q_{\text{n-gram}} \right|}$$

Reasoning Module

- **DeBERTa-V3-Large**
  - Has shown strong performance across a variety of tasks
  - DeBERTa-V3-Large pretrained on tasksource (520+ tasks) might be better (we leave this for future work)
- **Direct Training is Insufficient**
  - We anecdotally observed that directly training DeBERTa-V3-Large on QuALITY gave poor accuracies, due to the limited amount of training data (~38% dev)
- **RACE Pretraining**
  - RACE has 100k+ multiple choice questions based on over 28,000 passages, from the Chinese Middle and High School English exams
  - We pretrain the DeBERTa-V3-Large on RACE until accuracy saturates before fine-tuning on QuALITY

Results

Method	Description	Accuracy
No Context	Only the question is provided as input	44.55
Random	Context is selected randomly	50.38
TF-IDF	TF-IDF similarity is used to select sentences	55.44
BM25	The Okapi BM25 [4] [17] is used to retrieve the sentences	52.91
GraphCut-MI	$I_f(A;Q) = 2\lambda \sum_{i \in A} \sum_{j \in Q} s_{ij}$	55.10
FL-MI	$I_f(A;Q) = \sum_{i \in V} \min(\max_{j \in A} s_{ij}, \eta \max_{j \in Q} s_{ij})$	50.24
LogDet-MI	$I_f(A;Q) = \log\det(S_A) - \log\det(S_A - \eta^2 S_{A,Q} S_Q^{-1} S_{A,Q}^T)$	54.55
F+LogDet-MI	Filtration followed by LogDet-MI	52.29
CoM-MI	$I_{f_\eta}(\mathcal{A};\mathcal{Q}) = \eta \sum_{i \in \mathcal{A}} \psi(\sum_{j \in \mathcal{Q}} s_{ij}) + \sum_{j \in \mathcal{Q}} \psi(\sum_{i \in \mathcal{A}} s_{ij})$	55.99
F+CoM-MI	Filtration followed by CoM-MI	56.14

Table 2: Results on the QuALITY dataset

Analysis

- **Why does random perform reasonably?**
  - *DeBERTa could have already "seen" the QuALITY articles during pretraining*
  - We also evaluate the selections w.r.t the concatenation of the question and correct answer in terms of ROUGE *recall*
    - Based on *n*-grams, effectively measures how well "keywords" were selected
    - **Graph Cut:** ROUGE-1 = `0.58`, ROUGE-2 = `0.11`, ROUGE-L = `0.44`
    - **Random:** ROUGE-1 = `0.53`, ROUGE-2 = `0.07`, ROUGE-L = `0.40`
    - **BM25:** ROUGE-1 = `0.65`, ROUGE-2 = `0.14`, ROUGE-L = `0.49` *(keywords aren't everything!)*

Conclusion

- We propose a submodular conditional context selection module to select the sentences relevant for answering a given question
- We evaluate, compare, and analyze various submodular mutual information functions for this task
- We show the competitiveness of our method by evaluating it on the challenging QuALITY dataset (Pang et. al.)
  - Current best model (CoLISA) achieves 62.3% on the real QuALITY test set, next best is 55.4%
    - CoLISA uses contrastive learning, orthogonal to our modifications
  - Our best (but on our test split) was 56.14%
- In the future we plan to:
  - Evaluate our methods on more datasets from the SCROLLS benchmark (Shaham et. al.)
  - Experiment with more submodular functions outside of `submodlib`

Thank You

Extra Slides