AnswerSumm: A Manually-Curated Dataset and Automatic Pipeline for Answer Summarization

Alexander R. Fabbri[1], Xiaojian Wu[2], Srini Iyer[2], Haoran Li[2], Mona Diab[2]
[1]Department of Computer Science, Yale University [2]Facebook AI

Introduction

Community Question Answering (CQA) forums such as Stack Overflow and Yahoo! Answers contain a rich resource of answers to a wide range of questions. Each question thread can receive a large number of answers with different perspectives.

A major obstacle for multi-perspective, abstractive answer summarization is the absence of a dataset to provide supervision for producing such summaries. Recent work proposes heuristics to create data that are often noisy and do not cover all the perspectives present in the answers.

Contributions:

1) This work introduces a novel dataset of 4631 data points for answer summarization curated by professional linguists. Our annotation pipeline consists of sentence relevance, clustering, cluster summarization, and overall answer summarization to provide a comprehensive dataset of all subtasks in answer summarization.

2) We analyze and benchmark state-of-the-art models on the answer summarization task and demonstrate how data augmentation via a novel automatic dataset creation method further boosts automatic summarization performance by over 1 ROUGE-1 point.

Data Annotation

Data annotated by 10 professional linguists
Tasks:

- Relevance classification - label answer sentences as relevant or not
- Clustering - group similar relevant sentences
- Cluster summarization - summarize individual clusters
- Fuse cluster summaries

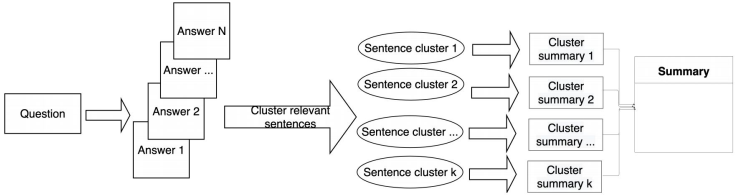

Figure 1: An illustration of our manually-curated dataset. Given a question and answers, professional linguists assign relevance labels, cluster those selected sentences, summarize each cluster and then fuse clusters into an overal summary.

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.

Table 1: An example summary from our answer summarization dataset, illustrating the multiple viewpoints present in the summaries created through our pipeline, and a subset of the 8 user answers to which the target summary can be aligned.

	True Rel	**True Not Rel**
Predicted Rel	4324	25088
Predicted Not Rel	5664	3349

Table 2: Confusion Matrix for RoBERTa relevance classification.

Task	**ROUGE-1/2/L**
Cluster Summarization	30.98/10.61/26.22
Cluster Fusion	51.64/32.67/47.13

Table 3: ROUGE scores for cluster summarization and cluster fusion tasks summarization tasks, showing cluster summarization as one of the bottlenecks in overall model summarization performance.

Model	**ROUGE-1/2/L**
BART-large (Lewis et al., 2020)	27.33/8.36/23.04
BART-large-aug (Lewis et al., 2020)	**28.92/9.09/24.18**
Pegasus-large (Zhang et al., 2020)	25.93/7.80/22.06
T5-base (Raffel et al., 2019)	24.62/7.19/20.73
BART-rel-oracle (Lewis et al., 2020)	**30.98/10.61/26.22**

Table 4: Model comparison for overal answer summarization task.

Data Statistics and Characteristics

- 4600+ data points annotated
- Input length ~600 tokens, summary length ~50 tokens
- Extractive ROUGE oracle: 40.05/18.45/35.70
- % novel unigrams: 20 (fairly extractive)
- All information in summaries is contained in the input (unlike XSum)

Modeling and Results

- Relevance classification: RoBERTa 0.49 F1 shows room for improvement and subjectivity of relevance labeling (Table 2)
- Cluster summarization is a difficult task due to compression and abstraction, but summary fusion is simple (Table 3)
- BART outperforms Pegasus and T5 on this dataset, and data augmentation via an automatic pipeline for creating heuristic summaries boosts performance.
- Much room for improvement

Conclusion

We propose multi-perspective answer summarization by introducing a manually-curated dataset for answer summarization created by professional linguists. We also propose a pipeline for data augmentation which mirrors the characteristics of our manual data and helps generalizability to new domains. Our dataset will promote work in all subtasks of answer summarization.