Group 101

Project 8

Healthcare Answer Summarization

INDRAPRASTHA INSTITUTE *of* INFORMATION TECHNOLOGY
DELHI

Shubhankar Tiwary	MT24139
Sai Krishna Kota	MT24078
Avikalp Rewatkar	MT24022

Introduction

- Healthcare platforms contain a vast amount of user-generated content: medical information, personal experiences, and recommendations.
- It makes difficult for users to extract important and relevant information regarding their question and answers.
- Our project helps users by generating summaries of answers for required perspectives.

PUMA Dataset

- The PUMA dataset is a collection of healthcare question-answer threads, annotated with diverse perspectives (e.g., INFORMATION, SUGGESTION, EXPERIENCE).
- The PUMA dataset includes 2,236 training samples, 959 validation samples, and 640 test samples.
- The PUMA dataset includes the following columns: uri, question, context, answers, labelled_answer_spans, labelled_summaries, and raw_text.

What we did!

<u>Baselines</u>

- Baseline 1 : Benchmarked Deepseek and BART with zero-shot setup.
- Baseline 2 : Replicated PLASMA model as guided in the paper.

Model	BERT-F1	BLEU-4
BART	0.828	0.109
DeepSeek	0.852	0.073

PLASMA Model	BERT-F1	BLEU-4
Paper	0.869	0.040
Our Model	0.810	0.007

What we did! (contd..)

Fine-Tuning

- BART : Used BART-Large-CNN model (406M parameters) to fine-tune on the given PUMA dataset .
- Flan-T5 : Used Flan-T5-Small model (77M parameters) to fine-tune on the given PUMA dataset.

Model	BERT-F1	BLEU-4
Flan-T5-Small	0.870	0.051
BART-Large-CNN	0.684	0.038

What we did! (contd..)

Mixture of Experts Architecture

We tried a MoE architecture by fine-tuning a separate model for each perspective of the summary.

Generated an expanded dataset where each instance has a unique perspective summary instead of having a dict of summaries and tuned 1 model for each perspective summary.

Then we created a pipeline to generate summary for respective perspective from that model.

What we did! (contd..)

Mixture of Experts Architecture

We used the following models :

- Flan-T5 : Fine-Tuned Flan-T5-base model for upto 3 epochs on the dataset , we also tried to set minimum length for generated summary.
- Flan-T5 with LoRA: Used Low Rank Adaptation technique to modify the fine-tuning method

Model	BERT-F1	BLEU-4
Overall	0.197	0.79
Information_Summary	0.166	0.50
Suggestion_Summary	0.199	0.38
Cause_Summary	0.256	1.68
Experience_Summary	0.234	3.38

Table 5: Evaluation metrics for different models of Flan-T5 with MoE Architecture

Model	BERT-F1	BLEU-4
Overall	0.192	0.50
Information_Summary	0.150	0.24
Suggestion_Summary	0.210	1.60
Cause_Summary	0.248	0.23
Experience_Summary	0.228	2.73

Table 7: Evaluation metrics for Flan-T5 with LoRA on MoE Architecture

Conclusion

1. For Fine-tuning:

   We concluded that that standard fine-tuning of Flan-T5-Small performs better than BART-Large-CNN with comparatively high scores.

2. For Mixture of Experts (MoE):

   The Mixture of Experts architecture shows potential for perspective-aware summarization but is limited by insufficient and imbalanced training data, especially for under-represented perspectives (such as QUESTION perspective).

Thank You

Group 101