Exploring Prompt Engineering for Few-Shot Text Summarization

Fawaizat Ahmed, Marcus Dahl, Amir Khodabakhshi, Saeed Mashhadi, Lei Meng

What have you done in this project?

- Emerging Abilities
- Text Summarization
- General and State-Of-The-Art models
- Evaluation Metric
- Human Evaluation

Zero and Few Shot prompting

Our main question on this project was

¨How does prompt engineering affect the performance of text summarization models?¨

Specifically:

Do general text generation models such as llama produce better or biased summaries than summarization based models such as Bart,e.t.c.

Does few-shot prompting improve the quality of model generated summaries compared to zero shot prompting.

How does different datasets such as professionally written papers vs informal blog post impact odel performance.

How does Rouge-based evaluation compare to human evaluation in assessing summaries.

Zero and Few Shot prompting

- Zero-Shot

  Zero Shot prompting means providing the model with a of direct instruction to perform a task without including any example demonstrations within the prompt.

  - How this works:
    - The model relies entirely on its pretrained knowledge and internal representations to generate the summary
- Few-Shot

  Few-shot prompting involves providing the model with a set of examples(2-3 but up to 7-8 for very LM) of the task along with their desired outputs. These examples are put into the prompt.

  - How this works:
    - we include a set of examples pairs: an input like a new and the corresponding desired output , then append the new input text for which we need the model to summarizes.

Models & Datasets

- Models:
  - Llama 3.1.3B & 3.1.1B
  - Pegasus
  - BART
- Datasets
  - Reddit TIFU
    - Informal stories, with human written “tl;dr”.
  - CNN/Dailymail
    - News articles with human written “highlights”.

Rouge-score and human evaluation

- What is rouge score?
  - Compares overlap of words (n-grams) between model output and reference summary.
  - 0-1, Higher the better.
  - Limitations include:
    - Doesn’t take grammar or synonyms into account.
- Small human evaluation (3 people)
  - Saw text and reference, was asked to rank generated summaries from best to worst.
  - Participants preferred zero-shot summaries to few-shot summaries.
  - Models possibly confused by few-shot prompting.
    - In one case the model summarized one of the examples.

Results rouge-scores

CNN/dailymail	Rouge-1	Rouge-2
Pegasus (from paper)	0.439	0.2120
Llama 1 (0shot)	0.3193	0.1138
Llama 1 (3-shot)	0.2866	0.0898
Llama 3 (0-shot)	0.2920	0.0979
Llama 3 (3-shot)	0.3552	0.14
BART (0-shot)	0.4087	0.1752

Reddit TIFU	Rouge-1	Rouge-2
Pegasus (from paper)	0.2654	0.0894
Llama 1 (0shot)	0.1016	0.0409
Llama 1 (3-shot)	0.1963	0.0265
Llama 3 (0-shot)	0.1789	0.0123
Llama 3 (3-shot)	0.2993	0.0741
BART (0-shot)	0.1586	0.0201

Conclusions

- Unclear if three-shot prompting beneficial in our experiments
    - Looking at some outputs, it does seem to help generate desired formats (e.g. length of summary)
    - ...But can also lead to confusion, as seen in at least one other paper [1]
    - Longer summaries can lead to higher rouge
        - Other rouge metrics exist, we took what seemed most common in research
- 3-shot is slower than 0-shot
- What about more than 3-shot? 10-shot? 100-shot?
    - Research indicates there is a point of diminishing returns [1]

[1] Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm https://dl.acm.org/doi/pdf/10.1145/3411763.3451760

Questions?