Examining Extractive and Abstractive Question Answering Methods for Climate Texts

Marisa Lenci and Grace Murnaghan
Spring 2025

Overview

Dataset: Climate Change Machine Reading Comprehension as introduced by Rony, et al. as part of Climate Bot (2022).

- 21,000 question-answer pairs each associated with one of 7,400 context passages
- Format is similar to SQuAD: labeled answers are spans of the text

Previous work:

- BERT models
  - Rony, et al., 2022 – create a “retriever” and a “reader” to find appropriate context given a question and to return an excerpt that answers the question, given a corpus of climate-related passages, using extractive approach
  - ClimateBERT (Webersinke, et al., 2022) – additional pre-training on top of DistilRoBERTa for climate texts
- LLMs
  - Nguyen, et al. (2024) compare various LLMs for question answering on climate texts

Our task: Build a question answering system specifically in the climate science domain. Compare extractive and abstractive methods across task- and domain-specific models. Evaluate models with context given in the dataset as well as with an information retrieval component.

Example QA pairs

Specificity is easier for QA and retrieval

Question: When did Lost Creek wildfire occur?

Context: here, impairment of water quality by wildfires in forested source water regions was examined as a critical vulnerability of downstream water treatment processes. in 2003, one of the most severe recorded fires (lost creek wildfire) occurred in the eastern slopes of the rocky mountains of southern alberta, canada and impacted several aspects of water quality and streamflow in the upper oldman river basin (orb). …

Label Answer: in 2003

Vague questions are challenging

Question: What it would take?

Context: there is a need for careful case studies of what adaptation would involve in particular locations, and what component of this is from infrastructure. …

Label Answer: it would take only a few such studies of major cities particularly at risk from climate change and with large infrastructure deficits to show that the unfccc estimates for africa and for 'developing asia' are far too low

Experiments

Approach:

1. Establish baseline
    a. Extractive: BERT
    b. Abstractive: T5
2. Compare to alternative pre-trained models out-of-the-box
    a. Extractive: DistilRoBERTa, ClimateBERT, BART for SQuAD, BART for Climate
    b. Abstractive: Flan-T5, Flan-T5 for Climate QLoRA, Llama
3. Fine tune pre-trained models, compare
    a. Extractive: DistilRoBERTa, ClimateBERT, BART for SQuAD, BART for Climate
    b. Abstractive: T5, Flan-T5 for Climate QLoRA, Llama
4. Further fine tune best models
    a. Extractive: BART for SQuAD
    b. Abstractive: T5

Evaluation: ROUGE, semantic similarity

Interesting Findings Along the Way

- Models fine tuned on climate-related texts often did not improve upon the general-purpose versions
- We had mixed results with fine tuning extractive models for our task
  - Had success with fine tuning distillroberta-base as measured by ROUGE
  - Bart-large-cnn-climate-change-summarization: training QA head for additional epochs or more layers using LoRA had small impact on ROUGE scores
    - We may have needed to train for longer to see benefit
- We had better results with fine tuning models that already worked with the structure of our data
  - Bart-large-finetuned-squadv1: best performing extractive method after fine tuning with LoRA
  - T5-small: best performing abstractive method after fine tuning

What Made This Hard / Things We Didn’t Expect

- Working with limited compute, GPU resources
- Dataset structure and style
  - More like factual, reading comprehension style question answer pairs based on context paragraphs than understanding conceptual climate science
  - 3 question answer pairs : 1 context paragraph, so no question answer pairs referred to multiple context paragraphs
- Simple tf-idf retrieval outperformed embedding-based retrieval

Results - Extractive Methods

Model	ROUGE-1	ROUGE-2	ROUGE-L
Baseline: bert-base-cased-squad2	0.27399	0.23738	0.27115
distilroberta-base	0.20000	0.13269	0.16757
distilroberta-base (train QA head for 3 epochs)	0.34177	0.22910	0.28975
distilroberta-base-climate-f	0.18772	0.13806	0.16642
distilroberta-base-climate-f (train QA head for 3 epochs)	0.12461	0.07747	0.10801
bart-large-finetuned-squadv1	0.43411	0.37962	**0.43218**
bart-large-finetuned-squadv1 (fine tuned with Lora for 2 epochs, r=8, alpha=32)	0.44285	0.36825	0.40077
bart-large-finetuned-squadv1 (fine tuned with Lora for 2 epochs, r=8, alpha=16)	**0.46334**	**0.39130**	0.42004
bart-large-finetuned-squadv1 fine tuned and run with tf-idf context retrieval	*0.35688*	*0.27123*	*0.31737*
bart-large-cnn-climate-change-summarization (train QA head for 2 epochs)	0.32481	0.23233	0.28355
bart-large-cnn-climate-change-summarization (train QA head for 5 epochs)	0.32636	0.23451	0.28362
bart-large-cnn-climate-change-summarization (train using LORA for 2 epochs)	0.32957	0.22386	0.28006

Results - Abstractive Methods

Model	ROUGE-1	ROUGE-2	ROUGE-L	Semantic Similarity
Baseline: T5-base	0.39642	0.33858	0.39496	0.5784
T5-small (fine tuned for 1 epoch)	0.57041	0.52688	0.56100	0.7301
T5-small fine tuned and run with tf-idf context retrieval	*0.41986*	*0.35099*	*0.40546*	*0.5813*
T5-small (fine tuned for 3 epochs)	**0.57813**	**0.53708**	**0.57066**	0.7341
T5-small fine tuned and run with tf-idf context retrieval	*0.42608*	*0.35809*	*0.41258*	*0.5852*
flan-t5-base	0.34952	0.28607	0.34610	0.5487
flan-t5-climate-qlora	0.34902	0.28667	0.34593	0.5487
Meta-Llama-3.1-8B-Instruct (temperature=1, top_p=0.3)	0.53869	0.44541	0.49941	**0.7679**
Llama-3.1-8B-Instruct run with tf-idf context retrieval	*0.42384*	*0.32620*	*0.38713*	*0.6696*

Analysis & Takeaways

- Abstractive methods performed better than extractive on ROUGE and semantic similarity
  - These results were also more human readable
- Task-specific pre-training seemed more important than domain-specific for our dataset
  - Models pre-trained for our task but not our domain performed better than models pre-trained for our domain but not our task
  - Fine tuning for our dataset on task-specific pre-trained models showed stronger impact
  - This could also be due to the dataset structure and style being reasonably well represented in general texts and not needing climate terminology to perform well, and reading comprehension style question answer pairs instead of conceptual understanding of climate science