Deep Climate Change: A Dataset and Adaptive domain pre-trained Language Models for Climate Change Related Task

Saeid A. Vaghefi, Veruska Muccione,
Christian Huggel, Hamed Khashehchi,
Markus Leippold
University of Zürich

NEURAL INFORMATION PROCESSING SYSTEMS

We propose climateGPT2 models, transformer-based language models that are further pretrained abstracts of climate-related articles

We collected a climate change corpus consisting of over 360 thousands of abstracts of top climate scientists' articles from trustable sources covering large temporal and spatial scales.
Comparison of the performance of GPT2 LM and our 'climateGPT2 models', fine-tuned on the CC-related corpus, on claim generation (text generation) and fact-checking, downstream tasks show the better performance of the climateGPT2 models compared to the GPT2.

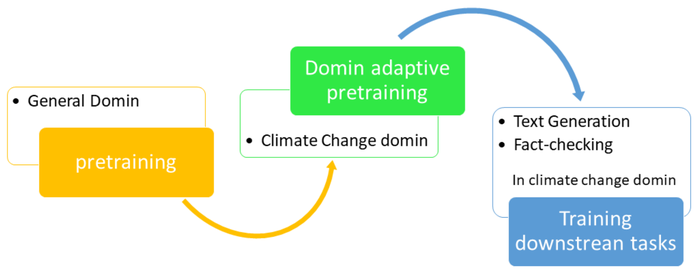
- climateGPT2 models decrease validation loss to 1.08 for claim generation from 43.4 obtained by GPT2.
- climateGPT2 models improved the masked language model objective for the fact-checking task by increasing the F1 score from 0.67 to 0.72

Contribution and Method

Contribution

The contributions of our paper are as follows: 1) 'climateGPT-2 models' are the first domain-specific GPT-based models pre-trained on climate change corpora for 3 days on two NVIDIA Gforce 2800 GPUs. 2) We show that pre-training GPT-2 on climate change corpus improves its performance on two downstream tasks: text generation (claim generation) and fact-checking. 3) We make our pre-processed datasets, the pre-trained weights of climateGPT-2, and the source code for fine-tuning climateGPT-2 models publicly available

Method



Climate domain corpus

We collected a large corpus of text from the publication of well-known scientists in the climate field. Therefore, we used the 1000 hot scientists list of climate change published by the Reuters press

Corpus	Num of paragraphs	Domain	Reference
climateGPT-2	360,233	Climate Change	https://www.reuters.com/investigates/special-report/climate-change-scientists-list

Fact-checking dataset

We used the Climate Fever dataset for the fact-checking task Diggelmann et al. (2020). Climate Fever consists of roughly 1,500 claims in the climate domain. Annotators have classified claims in the climate fever dataset as supported, refuted, or not enough by evidence sentences

Results

Text Generation (Climate Change claims

Loss of GPT-2 vs climateGPT-s models on the CC-related corpus

Model	Val Loss
GPT-2	43.45
climateGPT-2 (cGPT-2)	1.08
climateGPT-2 (ckeyGPT-2)	1.56

Fact-checking

The F1 score results show that climateGPT-2 models outperform GPT-2 by lowering F1 from 1.17 to 0.8.
By comparing our results with the fact-checking results performed on climate fever data with climateBERT (Webersinke et al. (2021), we realized that our results align with that study. However, our in-house corpus for adaptive pre-training is different from the climateBERT.

Results, average Validation Loss and average weighted F1 score on the fact-checking task on CLIMATE-FEVER dataset

Model	Val Loss	F1
GPT-2	1.17	0.67
climateGPT-2	0.83	0.72

climateGPT-2 models, Inputs/Outputs

Comparison of Climate Change related generated claims by GPT-2 vs climateGPT-2 models

Claim generation downstream task		
GPT-2	**c GPT-2**	**ckey GPT-2**
Given title: Climate change can affect hydropower operations through changes in the timing and magnitude of precipitation patterns **Given Keywords:** NA	**Given title:** Climate change can affect hydropower operations through changes in the timing and magnitude of precipitation patterns **Given Keywords:** NA	**Given title:** Climate change can affect hydropower operations through changes in the timing and magnitude of precipitation patterns **Given Keywords:** climate change', 'hydropower', 'energy', 'mitigation', 'adaptation
Generated Claims: Climate change can affect hydropower operations through changes in the timing and magnitude of precipitation patterns, including drought. However this may not be enough to offset all impacts on water resources that occur over a period or several decades due primarily for some reason from climate variability such as increases across regions within each state's watersheds	**Generated Claims:** Hydroelectricity generation is expected to be affected by climate variability, which may have significant impacts on water resources systems. This study evaluates possible implications for reservoir operation under changing hydroclimate conditions using a high-resolution global river discharge model driven with two regional models.	**Generated Claims:** Climate-induced shifts will alter energy systems throughout much higher latitudes as well; however climate mitigation is not always possible for some regions or even globally at present because there has been little work to date about how these responses vary across different levels within a given basin due primarily to human activities.

Conclusion

We showed that pre-training GPT-2 on climate change corpus largely improved its performance. climateGPT-2 obtains higher F1 scores in climate change fact-checking (0.72) compared to GPT-2 (0.67) and a lower validation loss (1.08) in climate change text generation task compared to the original GPT-2 (43.45)

References

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. arXiv preprint arXiv:2012.00614

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2021.https://doi.org/10.48550/ARXIV.2110.12010 Climatebert: A pretrained language model for climate-related text.

Contact

Saeid A. Vaghefi
University of Zürich
Switzerland
saeid.vaghefi@geo.uzh.ch