Automatic Explanation Generation For Climate Science Claims

Rui Xing[1] **Shraey Bhatia**[1] **Timothy Baldwin**[1,2] **Jey Han Lau**[1]
[1]The University of Melbourne [2]MBZUAI
ruixing@student.unimelb.edu.au, shraeybhatia@gmail.com,
tb@ldwin.net, jeyhan.lau@gmail.com

Abstract

Climate change is an existential threat to humanity, the proliferation of unsubstantiated claims relating to climate science is manipulating public perception, motivating the need for fact-checking in climate science. In this work, we draw on recent work that uses retrieval-augmented generation for veracity prediction and explanation generation, in framing explanation generation as a query-focused multi-document summarization task. We adapt PRIMERA to the climate science domain by adding additional global attention on claims. Through automatic evaluation and qualitative analysis, we demonstrate that our method is effective at generating explanations.

1 Introduction

The rapid dissemination of misinformation and disinformation through social media is a pressing issue, especially in the domain of climate science (Diggelmann et al., 2020; Anderegg et al., 2010) where climate change has become one of the biggest challenges to humankind. Claims such as *97% consensus on human-caused global warming has been disproven* seed scepticism, discredit climate science, and manipulate public perception and interpretation. To alleviate the influence of such potentially false claims, experts have increasingly engaged in science communication, including investigating such claims based on scientific evidence through websites such as climatefeedback.org and skepticalscience.com. This paper concerns the use of external knowledge to semi-automate the process of claim verification, as an assistive technology for contributors to such websites.

Inspired by recent work on retrieval-augmented generation (Lewis et al., 2020) and explainable fact-checking (Atanasova et al., 2020), we aim to (semi-)automate the process of claim veracity classification along with explanation generation. Our work draws on previous work on generating explanations in the climate science domain (Bhatia et al., 2021a) in using claims to retrieve relevant documents from knowledge sources and then generate explanations based on these documents. Unlike prior work, we frame it as a *query-focused summarization* task (Mollá et al., 2020; Sarker et al., 2013), where the query is a claim in our case, and the goal is to summarize information from the retrieved documents that addresses the claim. We evaluate our framework quantitatively and qualitatively, and explore the impact of different variants of attention on explanation generation.[1]

Text	**Label**
C: Sea-level rise is not accelerating.	REFU
E1: Climate-change driven accelerated sea-level rise detected in the altimeter era.	REFU
E2: Antarctica ice melt has accelerated by 280% in the last 4 decades.	REFU
E3: However scientists have found that ice is being lost, and at an accelerating rate.	REFU
E4: Climate scientists expect the rate to further accelerate during the 21st century.	NO_INFO
E5: More precise data gathered from satellite radar measurements reveal an accelerating rise of 7.5cm (3.0in) from 1993 to 2017, which is a trend of roughly 30cm (12in) per century.	NO_INFO

Table 1: An example claim ("C") and associated evidence passages ("Ek") from Climate-Fever ("REFU" = REFUTES; "NO_INFO" = NOT_ENOUGH_INFO).

2 Related Work

Fact checking is the task of assessing whether a textual claim is true, based on a corpus or knowledge base. Conventionally, the task is performed manually by human experts (Hassan et al., 2015). However, manual efforts do not easily scale (Elazar et al., 2021), leading to increasing attention in automatic fact checking (Wang, 2017; Alhindi et al.,

[1]The code associated with this paper is available at https://github.com/ruixing76/ClimateChange-ExpGen

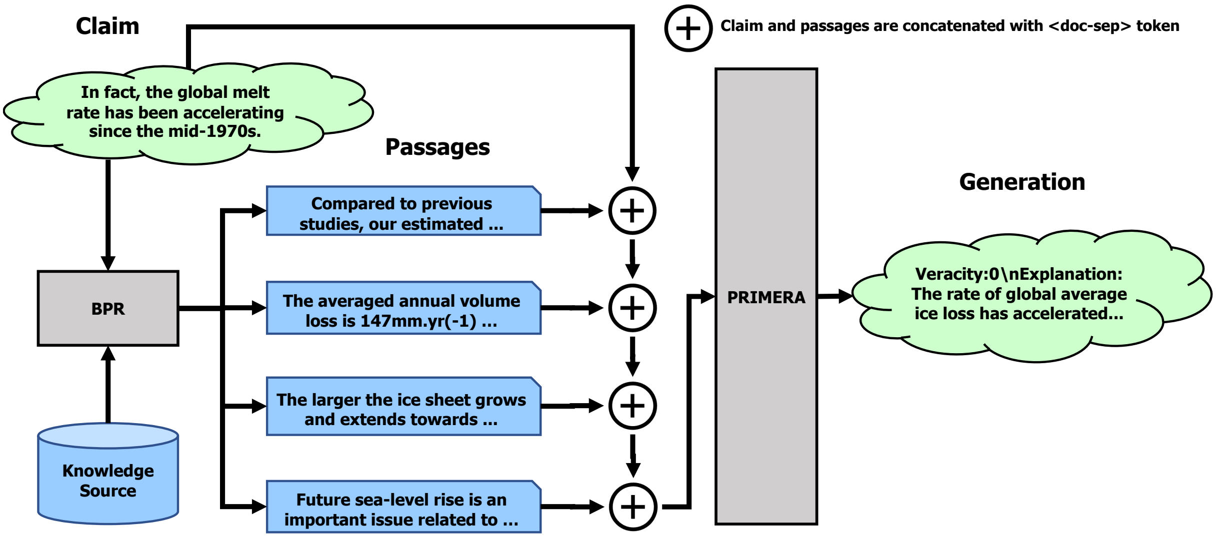

Figure 1: Overview of our method. First the claim is used as input to `BPR` to retrieve top-k claim-relevant passages (k is an adjustable hyperparameter, in this example k=4). Then the claim and passages are concatenated with `<doc-sep>` tokens for input to `PRIMERA`. Finally `PRIMERA` generates explanations together with veracity labels.

2018; Xu et al., 2019; Stammbach and Neumann, 2019; Atanasova et al., 2020). Debunking simply by assigning a *false* label to the claim is not persuasive, and can even reinforce mistaken beliefs (Lewandowsky et al., 2012). As such, it is necessary for automated fact-checking methods to provide explanations to support model predictions. For example, Popat et al. (2018) used attention-based methods to highlight salient excerpts from evidence articles, and Gad-Elrab et al. (2019) adopted knowledge bases to mine explanations. Atanasova et al. (2020) framed explanation generation as a joint classification and extractive summarization task. During generation, the model selects sentences from retrieved documents as explanations.

Separately, there has been recent work on extracting parameterized knowledge from large language models (Roberts et al., 2020), as well as augmenting them using external knowledge sources through retrieval augmentation (Karpukhin et al., 2020; Lewis et al., 2020; Yamada et al., 2021). Here, a claim or question is used to retrieve documents, which are fed into the generator as additional inputs, as a means of extending and domain-adapting large language models without additional pre-training.

There has also been recent work on the applications of NLP to the domain of climate science. Bhatia et al. (2021b) explored automatic classification of neutralization techniques in discourse relating to climate change/science. Diggelmann et al. (2020) introduced Climate-Fever as a novel dataset for veracity prediction. The closest work to our own is that on explanation generation by Bhatia et al. (2021a), which is based on fusion in decoder (Izacard and Grave, 2021), a sequence-to-sequence model that takes as input the claim and passages sourced through retrieval augmentation (Karpukhin et al., 2020; Yamada et al., 2021).

Unlike prior work, we first approach the task via multi-document summarization (Zhang et al., 2020a; Liu and Lapata, 2019; Liao et al., 2018), with a focus on the claim; as such, our approach can be interpreted as query-focused summarization. Specifically, we adopt `PRIMERA` (Xiao et al., 2022), a state-of-the-art pre-trained encoder–decoder model for multi-document summarization.

3 Data

There are two key data components in our task: (1) an external knowledge source from which we retrieve documents; and (2) paired claim–explanation data, to serve as the input (claim) and output (explanation).

For the external knowledge source, we use climate science-related abstracts from PubMed and reports from the Intergovernmental Panel on Climate Change ("IPCC"). IPCC reports are written by a mix of scientists, experts, and policy makers and provide scientific, technical, and socio-economic knowledge on climate change and options to mitigate its impacts. We sample climate science-related publications using MeSH descriptors.

Climate-Fever (Diggelmann et al., 2020) contains 1,535 claims relating to climate change. See Table 1 for an example, wherein each evidence item is labelled as SUPPORTS, REFUTES, or NOT_ENOUGH_INFO with respect to the claim. These are used to label each claim as SUPPORTS (= at least one evidence item is SUPPORTS and all others are NOT_ENOUGH_INFO), REFUTES (= at least one evidence item is REFUTES and all others are NOT_ENOUGH_INFO), NOT_ENOUGH_INFO (= all evidence items are NOT_ENOUGH_INFO), or DISPUTED (= a mixture of SUPPORTS and REFUTES evidence items). Each claim has multiple evidence items, and we create multiple claim–evidence instances for each *congruent* evidence item.[2] We discard DISPUTED claims in this work.

In our framework, the claim serves as the input for us to query the knowledge source to retrieve related documents, and the evidence constitutes the *explanation* that we want to generate as output.

4 Method

In Figure 1, we present an overview of our method, which is made up of two components: (1) a document retriever; and (2) a generator. Given a claim c_i, the retriever retrieves k passages $\{p_1, p_2, ..., p_k | c_i\}$ from the knowledge source, based on which the generator generates a veracity label y_i along with explanation e_i.[3] The generator is an encoder–decoder model which jointly processes the retrieved passages and claim in the form of an abstractive summarization model.

We adopt Binary Passage Retriever (BPR) (Yamada et al., 2021) as the retriever. BPR is a memory efficient version of dense passage retriever (Karpukhin et al., 2020). It first uses two independent BERT (Devlin et al., 2019) encoders to encode question and passages into continuous embeddings and then incorporates a hashing layer to reduce computational cost for similarity calculation. BPR is trained with a multi-task objective over two tasks: effective candidate generation based on binary codes and accurate reranking based on continuous vectors. We use the official release of BPR[4] which was pre-trained on Natural Questions (Kwiatkowski et al., 2019) without fine-tuning, and consider each claim as the query to retrieve top-*k* relevant passages from our knowledge source.

For the generator, we adopt PRIMERA (Xiao et al., 2022) to generate explanations, where the input is the claim concatenated with the top-*k* retrieved passages. PRIMERA is designed for multi-document summarization with Entity Pyramid Masking, a novel pre-training strategy to select and aggregate salient information from multiple documents. PRIMERA uses Longformer (Beltagy et al., 2020) as its encoder, and replaces standard full self-attention with sparse self-attention, i.e. it features a combination of local attention (self-attention between tokens in a narrow context window) and global attention (selected tokens that attends to all other words).

We structure the input by adding <doc-sep> (a special token denoting a document separator) between passages, and concatenating them with the claim with another <doc-sep> token. Moreover, we prepend claims and passages with the special prefix <CLAIM:> and <PASSAGE:> tokens respectively (to provide explicit indication of their functions). By default, PRIMERA assigns global attention only to <doc-sep> tokens. We extend this idea by adding extra global attention to the *claim words* and the two special prefix tokens (<CLAIM:> and <PASSAGE:>). This is to better focus the model on the claim. We also perform veracity prediction by generating veracity labels together with explanations, following Bhatia et al. (2021b). That is, the output takes the form of `Veracity:[lab]\nExplanation:[exp]`, where `[lab]` is the veracity label and `[exp]` is the generated explanation.

5 Experiments

As our baseline, we compare against Bhatia et al. (2021a) who use a retrieval-augmented generation framework to jointly perform veracity prediction and explanation generation using fusion in decoder (Izacard and Grave, 2021) and model it as question answering task. Note that in their approach a claim is concatenated with *each* passage and these claim–passage pairs are encoded separately — so as to reduce the computational overhead due to full self-attention — before they are fed to the decoder. Our approach, on the other hand, frames the task as query-focused multi-document summarization, and the use of PRIMERA means we can use the concatenated claim and all passages as input due

[2]Using Table 1 as an example, we would create 3 claim–evidence instances (the 4th and 5th evidence items are discarded as they have different labels to the claim).

[3]To clarify, the veracity label is the claim label and the explanation is an evidence in Climate-Fever (Table 1).

[4]https://github.com/studio-ousia/bpr

CLAIM: About 60% of the warming observed from 1970 to 2000 was very likely caused by the above natural 60-year climatic cycle during its warming phase. **LABEL:** `REFUTES` **GEN:** In the scientific literature, there is an overwhelming consensus that global surface temperatures have increased in recent decades and that the trend is caused mainly by human-induced emissions of greenhouse gases. **REF:** It is extremely likely (95-100% probability) that human influence was the dominant cause of global warming between 1951-2010.
CLAIM: That humans are causing the rise in atmospheric CO2 is confirmed by multiple isotopic analyses. **LABEL:** `SUPPORTS` **GEN:** Human activity since the Industrial Revolution has increased the amount of greenhouse gases in the atmosphere, leading to increased radiative forcing from CO2, methane, tropospheric ozone, CFCs, and nitrous oxide. **REF:** While CO2 absorption and release is always happening as a result of natural processes, the recent rise in CO2 levels in the atmosphere is known to be mainly due to human (anthropogenic) activity.

Table 2: Example generated explanations with `P-full`. CLAIM=claim text, LABEL=claim label, GEN=generated explanation, REF=reference explanation.

Model	B-Score	R-1	R-L	Accuracy
`FiD`	0.26	0.25	0.22	0.55
`P-claim`	0.29	0.29	0.24	0.56
`P-full`	**0.32**	**0.33**	**0.28**	**0.60**

Table 3: Explanation generation and veracity prediction performance: B-Score=BERTScore, R-1=ROUGE-1 and R-L=ROUGE-L.

to its sparse attention mechanism. To clarify, the main difference between our model and Bhatia et al. (2021a) lies in the generator, as both models use `BPR` as the retriever. In terms of evaluation metrics we use ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020b) for assessing generation quality, and accuracy for veracity prediction.

5.1 Overall performance

Table 3 shows the results for Bhatia et al. (2021a) (`FiD`) vs. two variants of our method: (1) `PRIMERA` that uses only claim as input (`P-claim`); and (2) `PRIMERA` that uses both claim and retrieved passages as input (`P-full`). `P-full` outperforms the baseline model by a margin of 5–6%, and `P-claim` also outperforms the baseline by a margin of 2–4%, indicating that `PRIMERA` is a better model for our task. Table 2 presents some explanations generated by `P-full`. We can see that these generated examples constitute plausible explanations.

5.2 Error Analyses

Automatic evaluation metrics have well-known limitations (Koto et al., 2022), so we additionally perform manual analysis of 50 samples from `P-full` that have the lowest ROUGE-1 scores. We categorize four types of errors including *claim-copying* (14%), *insufficient explanation* (56%), *hallucination* (10%), and *no errors* (20%) and present some examples in Table 4. *Claim-copying* refers to directly copying claims as explanations, which lacks explanatory power. *Insufficient explanations* contain relevant keywords but are overall implausible. *Hallucination* means generated explanations gives the impression of being fluent and natural but are unfaithful or nonsensical to the source documents. And lastly, *no errors* are situations where the explanations are good but due to low lexical overlap they receive low ROUGE scores.

Based on the proportion of these error categories, degenerative behaviour (*claim-copying* and *hallucination*) is low, and the main issue is the inability of our model to generate convincing explanations (*insufficient explanations*), which is a hard problem.

To better understand the underlying causes of *insufficient explanations*, we further analyse these examples. We hypothesise two possible causes: (1) the retrieved passages do not contain enough or relevant information (i.e. a retriever problem) and (2) the model is unable to reason or summarise over multiple passages (i.e. a generator problem). Our analysis reveals that the former is the main cause (21/28 samples), where the retrieved passages are not always relevant even though they share some keywords with the claim. This problem is exacerbated if the claim is complex and contains many important keywords, which makes the retrieved passages more noisy. Ultimately, our findings suggest that perhaps more focus should be put on improving the retriever, as it appears to be the main bottleneck in the system.

Error Type	Generation	R-1
Claim Copying	**CLAIM:** 'Global warming' is a myth — so say 80 graphs from 58 peer-reviewed scientific papers published in 2017. **LABEL:** `SUPPORTS` **GEN:** It is a myth — so say 80 graphs from 58 peer-reviewed scientific papers published in 2017. **REF:** Delingpole has engaged in climate change denialism; in 2009 he wrote of "The conspiracy behind the Anthropogenic Global Warming myth.	0.11
Insufficient Explanation	**CLAIM:** For instance, wind turbines kill birds and insects, and palm-oil plantations destroy the biodiversity of the rain forests. **LABEL:** `SUPPORTS` **GEN:** Tropical rain forests are known for their biodiversity and natural resources, and are considered key to the functioning of modern-day global climate. **REF:** Thousands of birds, including rare species, have been killed by the blades of wind turbines, though wind turbines contribute relatively insignificantly to anthropogenic avian mortality.	0.12
Hallucination	**CLAIM:** Climate change need not endanger anyone. **LABEL:** `REFUTES` **GEN:** Climate change does not endanger anyone, especially low-income people and those with limited economic resources. **REF:** Given the potential threat to marine ecosystems and its ensuing impact on human society and economy, especially as it acts in conjunction with anthropogenic global warming, there is an urgent need for immediate action.	0.11
No Error	**CLAIM:** Earth's changing climate is a critical issue and poses the risk of significant environmental, social and economic disruptions around the globe. **LABEL:** `SUPPORTS` **GEN:** Impacts include the direct effects of extreme weather, leading to injury and loss of life. **REF:** If no mitigating actions are taken, significant disruptions in the Earth's physical and ecological systems, social systems, security and human health are likely to occur.	0.13

Table 4: Error analysis on `P-full` CLAIM=claim text, LABEL=claim label, GEN=generated explanation, REF=reference explanation, R-1=ROUGE-1. R-1 is calculated between GEN and REF.

5.3 Analyzing different global attention

We next perform an ablation study with different forms of global attention in the encoder:[5]

- `P-full`: Our proposed model with global attention on special tokens and claim words.
- `−sep`: Global attention on claim words, special claim, and passage tokens only.
- `−claim`: Global attention on `<doc-sep>` only (default setting in Xiao et al. (2022)).
- `−all`: No global attention on any tokens (local attention only).

As shown in Table 5, `P-full` has the best performance. `−claim` has (marginally) lower performance than `−sep`, suggesting that the claim words are particularly important to the task. To better understand `P-full` vs. `−claim` (default `PRIMERA` configuration), we manually examine the quality of their generated explanations and observe that the latter is more likely to produce claim-copying errors and explanations that are inconsistent with the predicted veracity label. This indicates that the additional global attention helps the model to focus on claims to generate better and more consistent explanations.

[5]Note that sparse attention is only used for self-attention in the encoder; cross-attention from the decoder always uses full attention to the encoder inputs.

Setting	B-Score	R-1	R-L	Accuracy
`P-full`	0.31	0.33	0.28	0.60
`−sep`	0.30	0.33	0.28	0.57
`−claim`	0.29	0.31	0.26	0.59
`−all`	0.30	0.31	0.26	0.58

Table 5: Global attention results. B-Score=BERTScore, R-1=ROUGE-1 and R-L=ROUGE-L

6 Conclusion

In this work, we tackle the problem of claim veracity prediction and explanation generation in the domain of climate change. We use PubMed and IPCC reports as a knowledge source, and frame explanation generation as a query-focused summarization task and use `PRIMERA` as our generation model. Quantitative and qualitative analyses demonstrate that our proposed model improves the quality of generated explanations, and that additional global attention on the claim tokens is helpful.

References

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and Verification (FEVER)*, pages 85–90.

William R. L. Anderegg, James W. Prall, Jacob Harold, and Stephen H. Schneider. 2010. Expert credibility in climate change. *Proceedings of the National Academy of Sciences*, 107(27):12107–12109.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2021a. Automatic claim review for climate science via explanation generation. *CoRR*, abs/2107.14740.

Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2021b. Automatic classification of neutralization techniques in the narrative of climate change scepticism. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2167–2175, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. CLIMATE-FEVER: A dataset for verification of real-world climate claims. *CoRR*, abs/2012.00614.

Yanai Elazar, Hongming Zhang, Yoav Goldberg, and Dan Roth. 2021. Back to square one: Artifact detection, training and commonsense disentanglement in the winograd schema. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10486–10500. Association for Computational Linguistics.

Mohamed H. Gad-Elrab, Daria Stepanova, Jacopo Urbani, and Gerhard Weikum. 2019. Exfakt: A framework for explaining facts over knowledge graphs and text. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 87–95, New York, NY, USA. Association for Computing Machinery.

Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, page 1835–1838, New York, NY, USA. Association for Computing Machinery.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Fajri Koto, Timothy Baldwin, and Jey Han Lau. 2022. FFCI: A framework for interpretable automatic evaluation of summarization. *Journal of Artificial Intelligence Research*, 73.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3):106–131. PMID: 26173286.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Kexin Liao, Logan Lebanoff, and Fei Liu. 2018. Abstract Meaning Representation for multi-document summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1178–1190, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081, Florence, Italy. Association for Computational Linguistics.

Diego Mollá, Christopher Jones, and Vincent Nguyen. 2020. Query focused multi-document summarisation of biomedical texts. In *CLEF (Working Notes)*.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32. Association for Computational Linguistics.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Abeed Sarker, Diego Mollá, and Cécile Paris. 2013. An approach for query-focused text summarisation for evidence based medicine. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 295–304. Springer.

Dominik Stammbach and Guenter Neumann. 2019. Team DOMLIN: Exploiting evidence enhancement for the FEVER shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 105–109.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics.

Brian Xu, Mitra Mohtarami, and James Glass. 2019. Adversarial domain adaptation for stance detection. *arXiv preprint arXiv:1902.02401*.

Ikuya Yamada, Akari Asai, and Hannaneh Hajishirzi. 2021. Efficient passage retrieval with hashing for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 979–986, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

A Appendix

<s> CLAIM: 97 % consensus on human - ca used global warming has been dispro ven <doc-sep> PASSAGE: Since the mid - 19 th century , human activities have increased greenhouse gases such as carbon dioxide , methane , and nit rous oxide in the Earth 's atmosphere that resulted in increased average temperature . The effects of rising temperature include soil degradation , loss of productivity of agricultural land , desert ification , loss of biodiversity , degradation of ecosystems , reduced fresh - water resources , acid ification of the oceans , and the disruption and depletion of strat osp heric ozone . All these have an impact on human health , causing non - commun icable diseases such as injuries during natural disasters , malnutrition during famine , and increased mortality during heat waves due to complications in chronically ill patients . Direct exposure to <doc-sep> With a documented increase in average global surface temperatures of 0 . 6 degrees C since 1975 , Earth now appears to be warming due to a variety of clim atic effects , most notably the casc ading effects of greenhouse gas emissions resulting from human activities . There remains , however , no universal agreement on how rapidly , region ally , or asymm etr ically the planet will warm or on the true impact of global warming on natural disasters and public health outcomes . Most reports to date of the public health impact of global warming have been anecdotal and retrospective in design and have focused on the increase in heat - stroke deaths <doc-sep> Global air surface temperatures increased by about 0 . 6 degrees C during the 20 th century , but as Z w iers and Weaver discuss in their Perspective , the warming was not continuous . Two distinct periods of warming , from 1910 to 1945 and since 1976 , were separated by a period of very gradual cooling . The authors highlight the work by St ott et al ., who have performed the most comprehensive simulation of 20 th century climate to date . The agreement between observed and simulated temperature variations strongly suggests that forcing from anthrop ogenic activities , moder ated by variations in solar and volcanic forcing , has been the main driver of <doc-sep> Recent reconstruct ions of Northern Hemisphere temperatures and climate forcing over the past 1000 years allow the warming of the 20 th century to be placed within a historical context and various mechanisms of climate change to be tested . Compar isons of observations with simulations from an energy balance climate model indicate that as much as 41 to 64 % of pre anthrop ogenic (pre - 18 50) dec adal - scale temperature variations was due to changes in solar irrad iance and volcan ism . Removal of the forced response from reconstructed temperature time series yields residual s that show similar variability to those of control runs of coupled models , thereby lending support to the <doc-sep> The most pronounced warming in the historical global climate record prior to the recent warming occurred over the first half of the 20 th century and is known as the Early Tw ent ieth Century W arming (ET CW). Understanding this period and the subsequent slowdown of warming is key to dis ent angling the relationship between dec adal variability and the response to human influences in the present and future climate . This review discusses the observed changes during the E TC W and hypotheses for the underlying causes and mechanisms . Attribution studies estimate that about a half (40 - 54 %; p > . 8) of the global warming from 1901 to 1950 was <doc-sep> </s>

Figure 2: Visualization of attention weights on model input

A.1 Analyzing attention weights

Attention weights can provide insights into what the model focuses on during learning, and how it affects generation. We visualize attention strength on tokens in our model input in Figure 2. Darker shades indicate higher weights on corresponding words. We analyse the (summed) cross-attention weights on the input words at the final decoding step, and observe that our model tends to: (1) produce strong attention on the claim words and `<doc-sep>` tokens; and (2) focus on relevant words in the passages.

A.2 Implementation Details

We split Climate-Fever into training, validation and test sets which yields 963 training, 83 validation and 332 test instances. We trained `PRIMERA` with the following settings: number of retrieved passages = 5, batch size = 1 with gradient accumulation = 4, max input text length = 1,024 and max generated output length = 150. We use Adam optimizer, learning rate = 1e-5 with a linear scheduler, weight decay = 0.01, and total steps = 8,000 with warmup steps = 400. We evaluate our model on validation set every 500 steps. Following previous work (Bhatia et al., 2021a), we use ROUGE scores (ROUGE-1 and ROUGE-L) and rescaled BERTScore to evaluate the performance of explanation generation and classification accuracy (ACC) for veracity prediction.