Claim Verification via Knowledge-Grounding and Synthesizing Contrastive Arguments with Large Language Models

Xianyi Tuo (121090520)
School of Data Science
Chinese University of Hong Kong, Shenzhen
xianyituo@cuhk.edu.cn

Abstract

The spread of misinformation on digital platforms presents a significant challenge to information credibility and public trust. Traditional fact-checking methods are resource-intensive and struggle to keep up with the scale of online content, driving interest in automated claim verification systems. This paper proposes a novel approach combining retrieval-augmented generation (RAG) with contrastive argument synthesis to improve claim verification. Our method decomposes claims into sub-claims, retrieves external evidence using dynamic web searches, and generates both supporting and opposing arguments to assess claim veracity. This synthesis of contrasting arguments enhances accuracy and transparency in the verification process. We evaluate our approach on the HoVER dataset, showing strong performance in identifying "NOT_SUPPORTED" claims, with high recall. However, the system struggles with "SUPPORTED" claims, exhibiting low recall and precision.

1 Introduction

The proliferation of misinformation across digital platforms, especially social media, has emerged as a significant challenge, undermining public trust and the dissemination of reliable information[1]. Traditional fact-checking approaches, while effective, struggle to keep pace with the vast and rapidly evolving online content. Manual verification, though accurate, is inefficient and resource-intensive, especially in the face of the growing volume of claims that require scrutiny[2, 3]. As a result, there has been increasing interest in automated claim verification methods, leveraging the capabilities of large language models (LLMs) to efficiently handle this task at scale[4, 5, 6, 7].However, current approaches still face substantial hurdles. These include the need for large-scale, human-annotated datasets for training[8, 9], the risk of hallucination (where models generate plausible but inaccurate information)[10, 11], and the difficulty in grounding claims in reliable external sources[12, 13]. To address these challenges, recent work has explored retrieval-augmented generation (RAG) frameworks and symbolic reasoning techniques[14, 15], allowing models to access external, verifiable sources of information while maintaining interpretability and explainability in their decision-making processes[16, 17]. This report proposes a method inspired by the First-Order-Logic-Guided Knowledge-Grounded (FOLK) framework[15] and further enhanced by the idea of synthesis of contrastive arguments[18]. The approach leverages large language models to perform more accurate, dialectical claim verification by retrieving external evidence and generating both supporting and opposing arguments. By synthesizing contrastive arguments, the method aims to improve the accuracy of claim verification while providing transparent justifications for the model's decision-making process.

2 Related Work

Recent advances in large language models (LLMs) have significantly improved language understanding and generation, but they still face challenges such as hallucination, where models generate inaccurate information[19]. To address these limitations, retrieval-augmented generation (RAG) has been proposed, allowing models to access external, verifiable knowledge sources for more reliable outputs[20, 21]. While RAG has shown promise in many areas, its application to claim verification remains underexplored, particularly in ensuring accurate evidence retrieval and fine-grained classification of claims[15].

To improve claim verification, recent approaches have focused on grounding LLMs in external knowledge and synthesizing contrastive arguments, where models evaluate both supporting and opposing evidence to generate more nuanced veracity predictions[15, 18]. This work combines retrieval-augmented techniques with contrastive argument synthesis to enhance both the accuracy and interpretability of claim verification, offering a more robust solution to the challenge of misinformation detection.

3 Approach

Our approach builds upon the First-Order-Logic-Guided Knowledge-Grounded (FOLK) framework[15] but introduces several modifications to enhance claim verification. Like FOLK, our method begins with a decomposition of the claim into logical sub-claims, but instead of directly generating questions from predicates, we prompt the language model (LLM) to plan and structure the questions logically to gather more contextual information when searching. This enables the model to ask more domain specific questions, ensuring that the verification process considers all relevant details.

In the second stage, I implemented a ReAct[22] agent integrated with Google Custom Search Api. The agent autonomously queries the web to retrieve knowledge-grounded answers to the questions generated in the previous step. Unlike traditional fact-checking methods, our agent adapts its queries based on the domain-specific context of the claim, aiming to gather more relevant and precise information for claim verification. This process ensures that the model can dynamically search for additional data if needed.

The third stage involves the agent verifying each sub-claim by using the retrieved information, with the option to further query the web for additional context if required. The agent then attempts to prove the claim to be wrong or correct using evidence and reasoning of last two stages to generate contrastive arguments[18]. Finally, the agent makes a veracity prediction for the overall claim—either SUPPORTED or NOT_SUPPORTED—based on the gathered evidence and reasoning from question-answering, sub-claims verification and contrastive arguments. Each prediction is accompanied by a confidence score and a detailed explanation of the reasoning behind the decision for nuance inside the claim.

4 Experiments

We performed the experiment on part of the **HoVER** dataset[23]. Due to the free pricing limitation of Google Custom Search Api, This experiment only include part of the dataset

4.1 Data

HoVER [23] is a multi-hop fact verification dataset created to challenge models to verify complex claims against multiple information sources, or “hop”. This experiment directly use the processed one from FOLK with the validation set for evaluation since the test sets are not released publicly[15]. And this experiment only choose part of two hops part and three hops part

4.2 Evaluation method

This experiment use GPT-3.5-turbo as the underlying LLM and Coogle Custom Search Api as retrieval engine tool to obtain internet search result from the whole internet. For testing, we randomly chosed the data from the dataset to perform the evaluation

4.3 Experimental details

The detailed commands of setup is included inside the README file. The test data is randomly selected considering the free Google Custom Search Api may fail when running the program. For this experiment, 48 lines from two hops dataset and 45 lines from three hops dataset were tested.

4.4 Results

The model's performance across both hops shows modest accuracy, with 54.17% for hop two and 48.89% for hop three. It performs better in identifying the "NOT_SUPPORTED" class, with a recall of 0.80 in hop two and 0.78 in hop three. However, it struggles significantly with the "SUPPORTED" class, achieving a low recall of 0.26 in hop two and 0.18 in hop three. This results in poor F1-scores for "SUPPORTED" (0.35 for hop two and 0.26 for hop three). The confusion matrices reveal a high number of false positives for "SUPPORTED," indicating that the model is not effectively distinguishing between the two classes. The overall performance suggests that while the model can identify "NOT_SUPPORTED" instances fairly well, it needs further improvements, particularly in predicting "SUPPORTED" cases.

5 Analysis

In conclusion, the model performs reasonably well in identifying "NOT_SUPPORTED" instances but has difficulty accurately predicting the "SUPPORTED" class across both hops. The low recall for the "SUPPORTED" class and the relatively high number of false positives indicate that the model might benefit from improvements such as better class balancing, parameter tuning, or additional feature engineering to improve its ability to distinguish between the two classes. One possible reason could be the system still have shortcomings on the decomposition stage prompt design, as additional questions for context exploring always bring unrelated information which may confuse the verification process. Additionally, this system is still not robust enough as the LLM may not answer as the expected format. For example, this system detect special token '**END**' to end the ReAct process, but LLM may not output as expected and reach max turns and then agent will give "Sorry, I can not help you with that.".

6 Conclusion

In conclusion, this project demonstrates the potential of using retrieval-augmented generation (RAG) and contrastive argument synthesis for claim verification, achieving success in identifying "NOT_SUPPORTED" claims. However, the system struggles with accurately predicting "SUPPORTED" claims, due to issues with context retrieval and LLM output formatting. Key lessons include the importance of refining query design and improving the interaction between retrieval and reasoning. Future work could focus on more domain specific search tools, enhancing prompt design, and improving model output robustness to improve overall accuracy and reliability in real-world misinformation detection.

References

[1] Canyu Chen, Haoran Wang, Matthew Shapiro, Yunyu Xiao, Fei Wang, and Kai Shu. Combating health misinformation in social media: Characterization, detection, intervention, and open issues, 2022.

[2] Nicholas Micallef, Bing He, Srijan Kumar, Mustaque Ahamad, and Nasir Memon. The role of the crowd in countering misinformation: A case study of the covid-19 infodemic. In *2020 IEEE international Conference on big data (big data)*, pages 748–757. IEEE, 2020.

[3] Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. Automated fact-checking for assisting human fact-checkers. *arXiv preprint arXiv:2103.07769*, 2021.

[4] Kai Shu, Ahmadreza Mosallanezhad, and Huan Liu. Cross-domain fake news detection on social media: A context-aware adversarial approach. In *Frontiers in fake media generation and detection*, pages 215–232. Springer, 2022.

[5] Zhiwei Yang, Jing Ma, Hechang Chen, Hongzhan Lin, Ziyang Luo, and Yi Chang. A coarse-to-fine cascaded evidence-distillation neural network for explainable fake news detection. *arXiv preprint arXiv:2209.14642*, 2022.

[6] Zhenrui Yue, Huimin Zeng, Yang Zhang, Lanyu Shang, and Dong Wang. Metaadapt: Domain adaptive few-shot misinformation detection via meta learning. *arXiv preprint arXiv:2305.12692*, 2023.

[7] Eun Cheol Choi and Emilio Ferrara. Fact-gpt: Fact-checking augmentation via claim matching with llms. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 883–886, 2024.

[8] Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Gear: Graph-based evidence aggregating and reasoning for fact verification. *arXiv preprint arXiv:1908.01843*, 2019.

[9] Yiqiao Jin, Xiting Wang, Ruichao Yang, Yizhou Sun, Wei Wang, Hao Liao, and Xing Xie. Towards fine-grained reasoning for fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5746–5754, 2022.

[10] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[11] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[12] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*, 2020.

[13] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR, 2022.

[14] Xuan Zhang and Wei Gao. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. *arXiv preprint arXiv:2310.00305*, 2023.

[15] Haoran Wang and Kai Shu. Explainable claim verification via knowledge-grounded reasoning with large language models. *arXiv preprint arXiv:2310.05253*, 2023.

[16] Boshko Koloski, Timen Stepišnik Perdih, Marko Robnik-Šikonja, Senja Pollak, and Blaž Škrlj. Knowledge graph informed fake news classification via heterogeneous representation ensembles. *Neurocomputing*, 496:208–226, 2022.

[17] Ziyi Kou, Lanyu Shang, Yang Zhang, Zhenrui Yue, Huimin Zeng, and Dong Wang. Crowd, expert & ai: A human-ai interactive approach towards natural language explanation based covid-19 misinformation detection. In *IJCAI*, pages 5087–5093, 2022.

[18] Zhenrui Yue, Huimin Zeng, Lanyu Shang, Yifan Liu, Yang Zhang, and Dong Wang. Retrieval augmented fact verification by synthesizing contrastive arguments. *arXiv preprint arXiv:2406.09815*, 2024.

[19] Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*, 2023.

[20] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020.

[21] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[22] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023.

[23] Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. Hover: A dataset for many-hop fact extraction and claim verification. *arXiv preprint arXiv:2011.03088*, 2020.