Claim Decomposition for Fact Checking

Minghui Huang
minghuihuang@utexas.edu

Summary

For my master's thesis research, I will focus on evaluating the quality of claim decomposition in fact-checking processes and plan to present my findings at conferences like ACL.

Claim decomposition plays a crucial role in the fact-checking process. However, current studies mainly rely on prompted Large Language Model (LLM)-based methods to treat texts as claims and decompose them into atomic claims, with limited focus on assessing the quality of these decompositions. To address this gap, I propose to define three metrics - `COMPLETENESS`, `CORRECTNESS` and `SEMANTIC ENTROPY` - to automatically evaluate the quality of atomic claims produced by claim decomposition models. Using these metrics, I will train a lightweight claim decomposition model and optimize it by incorporating these metrics as the reward function. I will then evaluate the performance of this optimized model in downstream tasks. Additionally, by leveraging automatic evaluation, I plan to establish a benchmark for claim decomposition.

In future work, I plan to enhance the factual accuracy of responses generated by language models by integrating the atomic claims as a fact-datastore.

1 Introduction

Long-form text generated by large language models (LMs) has been widely used. However, these generated texts may contain factual errors. To address this issue, a lot of methods have been proposed to evaluate the factuality of the generated text. Current fact-checking methodologies, such as SAFE (Wei et al., 2024), FActScore (Min et al., 2023) and FacTool (Chern et al., 2023), inherently treat generated text as a claim and rely on claim decomposition as a foundational step in the fact-checking process.

Claim decomposition involves breaking down a claim—which could be a generated text or a split sentence within the generated text—into individual atomic claims (Wanner et al., 2024; Hu et al., 2024). Terms like ”atomic fact” and ”atomic proposition” are often used interchangeably with atomic claim. The fact-checking process generally entails decomposing the generated text into atomic claims, evaluating the factuality of each atomic claim using external knowledge sources, and then aggregating these results to determine the overall factuality of the generated text (Iqbal et al., 2024; Li et al., 2024). However, these approaches overlook the variability in the quality of claim decomposition itself. Since claim decomposition is a critical component of the fact-checking process, as it determines the number and scope of each evaluated atomic claim, the results of the analysis or the resulting metrics will inherently be influenced by the decomposition method (Hu et al., 2024). Therefore, a comprehensive evaluation of the decomposed claims is essential to ensure the accuracy and reliability of the fact-checking process.

Moreover, current methods for decomposing generated text by LMs into atomic claims rely on using the same LM for this task (Iqbal et al., 2024). However, this approach is problematic for several reasons: (1) coupling claim decomposition with the fact-checking of LM-generated text makes it difficult to determine whether issues arise from the LM's claim decomposition process or from the text generated by the LM itself; (2) automatic claim decomposition strategies using LMs are computationally expensive, requiring numerous calls to the model to evaluate a single response. Therefore, there is a need for a more easily evaluable, optimizable and lightweight claim decomposition model. My plan is to develop a small, high-performance, and cost-effective claim decomposition model to decouple the claim decomposition process from the LM, thereby addressing these challenges.

Three Metrics and A Claim Decomposition Model To develop an effective claim decomposition model, we must first define what constitutes a good atomic claim. We propose that a good atomic claim should exhibit three key characteristics: high completeness (Kamoi et al., 2023), high correctness (Wanner et al., 2024), and high semantic entropy (Farquhar et al., 2024).

- **High Completeness**: The decomposed claims should cover all necessary aspects of the original text.

- **High Correctness**: Each atomic claim should be correct, meaning it should be factual using the original text as evidence.

- **High Semantic Entropy**: Factual atomic claims should not be paraphrased repeatedly or exhibit semantic overlap with other factual atomic claims. All factual atomic claims should be highly atomic and non-overlapping in meaning, indicating that the semantic entropy of the decomposed claims should be high.

Based on these characteristics, we can derive three corresponding metrics to evaluate atomic claims automatically. With these metrics in place, we can establish an initial claim decomposition model and set optimization objectives. Subsequently, we can build a training process that employs reinforcement learning (Schulman et al., 2017; Rafailov et al., 2024), using these metrics as rewards to guide the generation of high-quality atomic claims.

Downstream Task Impact We also intend to conduct experiments to explore how the three metrics of claim decomposition correlate with downstream tasks.

A Benchmark Furthermore, we may try to extend these three metrics into building a benchmark for the claim decomposition task, such as (Zhang et al., 2024), enabling the evaluation of different models' performance on this task.

2 Related Works

Many existing works employ a pipeline approach to decompose text into atomic claims for evaluating the factuality of long-form text. Recent advancements in claim decomposition mainly rely on prompted LLM-based methods, often incorporating in-context example decompositions (Min et al., 2023; Kamoi et al., 2023; Chern et al., 2023; Wei et al., 2024; Iqbal et al., 2024). To leverage propositions, akin to atomic claims, as retrieval units, (Chen et al., 2024) developed a Propositionizer that decomposes text into simple propositions using data generated by GPT-4. (Song et al., 2024) introduced VeriScore to check whether each decomposed atomic claims are verifiable. (Gunjal and Durrett, 2024) decontext molecular facts to make each decomposed atomic claims can be self-contained. However, there has been limited attention to fully evaluate the quality of these decomposed atomic claims. (Wanner et al., 2024) highlighted that downstream fact-checking methods are sensitive to the decomposition approach and proposed ClaimScore to assess the number of supported atomic claims as a measure of decomposition quality. Nonetheless, this approach overlooks the semantic overlap between atomic claims and does not assess the completeness of the decomposed claims. If a language model repeatedly paraphrases factual claims, it may achieve a very high ClaimScore. Consequently, we plan to define metrics for automatically evaluating the quality of decomposed atomic claims and use these metrics as rewards to train a reinforcement-learning-based claim decomposition model. Our aim is to generate atomic claims that are complete (Kamoi et al., 2023), correct (Wanner et al., 2024), and exhibit high semantic entropy (Farquhar et al., 2024).

3 Methodology

Following the approach of FActScore (Min et al., 2023), we will first segment a text into sentences, treat each sentence as a claim, and then decompose it into atomic claims using a claim decomposition model to achieve full decomposition. We use $d(c) = \{ac_1, ..., ac_n\}$ to denote the automatic decomposition of a claim c into atomic claims $\{ac_1, ..., ac_n\}$ via the claim decomposition model $d(.)$.

3.1 Claim Decomposition Model

Evaluation Metrics To construct a robust claim decomposition model, we must first identify the objectives for training such a model. We will introduce COMPLETENESS, CORRECTNESS and SEMANTIC ENTROPY as our evaluation metrics.

- COMPLETENESS: After decomposing the original claim into atomic claims, we assess the completeness of these atomic claims in relation to the original claim. Completeness is measured by computing the semantic similarity between the decomposed atomic claims $\{ac_1, ..., ac_n\}$ and the original claim c using the formula $cp(c, ac) = P(y = equal|c, ac)$. The semantic similarity $P(y = equal|c, ac)$ can be determined using semantic similarity models, which adopt the similar architecture of SummaC (Laban et al., 2022). However, we need to construct synthetic training data for this semantic similarity model. Since atomic claims are distinct entities, while the original claim is a cohesive unit, they exist on different levels. To train such a model, we must develop a specialized training dataset.

- CORRECTNESS For each pair of decomposed atomic claim ac_i and original claim c, we infer the factual label {`entailment, neutral, contradiction`} using a natural language inference (NLI) model. The correctness is then evaluated by calculating the percentage of factual atomic claims that are labeled as {`entailment`} as $cr(c, ac) = \frac{1}{n}\sum_{i=1}^{n} NLI(c, ac_i) = entailment.$

- SEMANTIC ENTROPY: Factual atomic claims $\{ac_1, ..., ac_n\}$ can be grouped into clusters by determining whether they entail each other using a natural language inference (NLI) model. If $NLI(ac_i, ac_j) = entailment$ or $NLI(ac_j, ac_i) = entailment$, it indicates that ac_i and ac_j belong to the same cluster C. The semantic entropy is computed as $se(\{ac_1, ..., ac_n\}) = -\sum_C P(C|\{ac_1, ..., ac_n\}) \log P(C|\{ac_1, ..., ac_n\})$. This metric encourages the model to generate highly atomic (numerous) and non-overlapping (less clustered) atomic claims.

Both CORRECTNESS and SEMANTIC ENTROPY employ NLI models (Bowman et al., 2015; Williams et al., 2018; Nie et al., 2020; Liu et al., 2022), but require fine-tuning with training data at different granularities. For the CORRECTNESS metric, the model assesses whether an atomic claim, treated as a hypothesis, is true (entailment), false (contradiction), or undetermined (neutral) given a sentence as the premise. For the SEMANTIC ENTROPY metric, the model evaluates whether two decomposed atomic claims entail each other by treating each as both hypothesis and premise. Therefore, it is necessary to curate a tailored training dataset to ensure the model performs effectively in computing these metrics.

Training Pipeline It is hard to train a supervised claim decomposition model with annotated atomic claims to match these three metrics. But it is easy to use these metrics as rewards in a reinforcement-learning based training process to guide the claim decomposition model. We follow a pipeline similar to Reinforcement Learning from Human Feedback (RLHF) (Ziegler et al., 2020), which consists of three phases:

- Supervised Fine-Tuning (SFT): We start by training a claim decomposition model using supervised learning on high-quality data for claim decomposition.
- Reward Computation: We define the reward function as $\alpha * cp(c, ac) + \beta * cr(c, ac) + \gamma * se(ac)$, where α, β, and γ are hyperparameters.
- Reinforcement Learning (RL) Optimization: During the RL phase, the learned reward function is employed to provide feedback, which is then used to fine-tune the claim decomposition model.

This approach ensures that the model learns to generate high-quality atomic claims that are complete, correct, and exhibit high semantic entropy.

3.2 Downstream Tasks Evaluation

Conduct an ablation study to evaluate the impact of different components of the claim decomposition model on downstream fact-checking tasks. This study will help identify which metrics and training strategies contribute most significantly to improved fact-checking performance.

4 Research Plan

4.1 Propsed Timeline

- **Paper Submission**: Before January 15, 2026
- **Thesis Submission**: Before November 15, 2026

4.2 Ideas to Explore

Fact-Datastore

- *Objective*: Enhance factual response generation by integrating atomic claims.
- *Components*:
  - Experiment with Nonparametric LMs (Khandelwal et al., 2020; Shi et al., 2024; Min et al., 2024) using various granularities of context as the fact-datastore.
  - Employ atomic claims to weight the significance of each token, thereby refining evidence retrieval and enabling Nonparametric LMs to produce more reliable and factual responses.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. 2024. Dense X Retrieval: What Retrieval Granularity Should We Use? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15159–15177, Miami, Florida, USA. Association for Computational Linguistics.

I.-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. FacTool: Factuality Detection in Generative AI – A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios. *arXiv preprint*. ArXiv:2307.13528 [cs].

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Anisha Gunjal and Greg Durrett. 2024. Molecular Facts: Desiderata for Decontextualization in LLM Fact Verification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3751–3768, Miami, Florida, USA. Association for Computational Linguistics.

Qisheng Hu, Quanyu Long, and Wenya Wang. 2024. Decomposition Dilemmas: Does Claim Decomposition Boost or Burden Fact-Checking Performance? *arXiv preprint*. ArXiv:2411.02400 [cs].

Hasan Iqbal, Yuxia Wang, Minghan Wang, Georgi Nenkov Georgiev, Jiahui Geng, Iryna Gurevych, and Preslav Nakov. 2024. OpenFactCheck: A Unified Framework for Factuality Evaluation of LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 219–229, Miami, Florida, USA. Association for Computational Linguistics.

Ryo Kamoi, Tanya Goyal, Juan Rodriguez, and Greg Durrett. 2023. WiCE: Real-World Entailment for Claims in Wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583, Singapore. Association for Computational Linguistics.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through Memorization: Nearest Neighbor Language Models. In

8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Haonan Li, Xudong Han, Hao Wang, Yuxia Wang, Minghan Wang, Rui Xing, Yilin Geng, Zenan Zhai, Preslav Nakov, and Timothy Baldwin. 2024. Loki: An Open-Source Tool for Fact Verification. *arXiv preprint*. ArXiv:2410.01794 [cs].

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI Collaboration for Natural Language Inference Dataset Creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sewon Min, Suchin Gururangan, Eric Wallace, Weijia Shi, Hannaneh Hajishirzi, Noah A. Smith, and Luke Zettlemoyer. 2024. SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 53728–53741, Red Hook, NY, USA. Curran Associates Inc.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint*. ArXiv:1707.06347 [cs].

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. REPLUG: Retrieval-Augmented Black-Box Language Models. In *Proceedings of the 2024 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 8371–8384, Mexico City, Mexico. Association for Computational Linguistics.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. VeriScore: Evaluating the factuality of verifiable claims in long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9447–9474, Miami, Florida, USA. Association for Computational Linguistics.

Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. 2024. A Closer Look at Claim Decomposition. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 153–175, Mexico City, Mexico. Association for Computational Linguistics.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form factuality in large language models. *arXiv preprint*. ArXiv:2403.18802 [cs].

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Zhihao Zhang, Yixing Fan, Ruqing Zhang, and Jiafeng Guo. 2024. A Claim Decomposition Benchmark for Long-form Answer Verification. *arXiv preprint*. ArXiv:2410.12558 [cs].

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-Tuning Language Models from Human Preferences. *arXiv preprint*. ArXiv:1909.08593 [cs, stat].