Evidence Retrieval is almost All You Need for Fact Verification

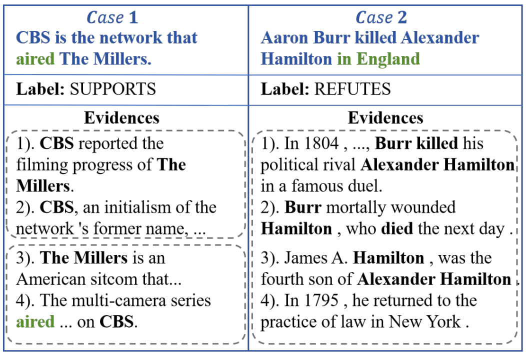
Liwen Zheng[1], Chaozhuo Li[1], Xi Zhang[1]*, Yuming Shang[1], Feiran Huang[2], Haoran Jia[1]
[1]Key Laboratory of Trustworthy Distributed Computing and Service (MoE),
Beijing University of Posts and Telecommunications, China
{zhenglw, lichaozhuo, zhangx, shangym, jiahaoran}@bupt.edu.cn
[2] College of Cyber Security/College of Information Science and Technology, Jinan University
huangfr@jnu.edu.cn

Abstract

Current fact verification methods generally follow the two-stage training paradigm: evidence retrieval and claim verification. While existing works focus on developing sophisticated claim verification modules, the fundamental importance of evidence retrieval is largely ignored. Existing approaches usually adopt the heuristic semantic similarity-based retrieval strategy, resulting in the task-irrelevant evidence and undesirable performance. In this paper, we concentrate on evidence retrieval and propose a **R**etrieval-**A**ugmented **V**erification framework RAV, consisting of two major modules: the hybrid evidence retrieval and the joint fact verification. Hybrid evidence retrieval module incorporates an efficient retriever for preliminary pruning of candidate evidence, succeeded by a ranker that generates more precise sorting results. Under this end-to-end training paradigm, gradients from the claim verification can be back-propagated to enhance evidence selection. Experimental results on FEVER dataset demonstrate the superiority of RAV.

1 Introduction

Fact verification endeavors to locate and incorporate credible evidence to autonomously assess the veracity of the target textual statements (Thorne et al., 2018). Existing works (Zhou et al., 2019; Liu et al., 2020; Wu et al., 2021) generally adhere to a two-stage learning paradigm: **evidence retrieval** to identify a set of key evidential sentences from a large corpus, and **claim verification** which determines the authenticity of a claim based on the semantic interactions between the claim and the retrieved evidence (Guo et al., 2022b). Considerable effort has been devoted to the claim verification stage (Zhong et al., 2020; Zou et al., 2023; Wu et al., 2021). However, solely advancing the claim verification might not be the panacea. The efficacy of verification is significantly contingent upon the quality of retrieved evidence (Hu et al., 2023). If the refereed evidence fails to provide accurate knowledge relevant to the claim, it would be intractable to achieve the correct decisions.



Case **1** **CBS is the network that aired The Millers.**	*Case* **2** **Aaron Burr killed Alexander Hamilton in England**
Label: SUPPORTS	**Label:** REFUTES
Evidences 1). **CBS** reported the filming progress of **The Millers**. 2). **CBS**, an initialism of the network 's former name, ... 3). **The Millers** is an American sitcom that... 4). The multi-camera series **aired** ... on **CBS**.	**Evidences** 1). In 1804 , ..., **Burr killed** his political rival **Alexander Hamilton** in a famous duel. 2). **Burr** mortally wounded **Hamilton** , who **died** the next day . 3). James A. **Hamilton** , was the fourth son of **Alexander Hamilton** . 4). In 1795 , he returned to the practice of law in New York .

Figure 1: The illustration of two FEVER cases.

Despite the significance of evidence retrieval, existing models (Zhou et al., 2019; Si et al., 2021) generally utilize a trivial retrieval strategy. All candidates are embedded into low-dimensional embeddings alongside the target claim. Subsequently, the k-nearest neighbor (KNN) search is employed to identify the top-k evidence based on their cosine similarity to the claim. However, such KNN-based strategies suffer from two limitations. Firstly, evidence that is most semantically similar to the claim may not be the most desirable for claim verification. Figure 1 illustrates two example claims along with their top-4 evidence retrieved by existing methods. Although semantically similar, most top candidates fail to provide clues to verify the statement in the claim. For example, while the first and second evidence of *Case 1* exhibits a high degree of semantic correlation with the claim, they lack the pivotal evidence of “aired”. Thus, simply selecting evidence based on heuristic semantic similarity may not be the optimal solution (Zhao et al., 2023). Secondly, evidence retrieval is independent from the claim verification, resulting in a separated training

[*] Correspondence to: Xi Zhang <zhangx@bupt.edu.cn>.

Findings of the Association for Computational Linguistics ACL 2024, pages 9274–9281
August 11-16, 2024 ©2024 Association for Computational Linguistics

paradigm. The training signals from the verification loss cannot guide the update of text encoder used in the evidence retrieval. Thus, the embeddings utilized for evidence retrieval may be task-irrelevant, resulting in undesirable performance.

In this paper, we concentrate on the crucial yet often overlooked evidence retrieval stage, with the goal of introducing a novel end-to-end training paradigm. Gradients from the claim verification could be back-propagated to the evidence retrieval stage to enhance evidence selection. In this manner, retrieved evidence is ensured to be task-relevant rather than merely similar to the claim. A straightforward strategy involves concatenating the claim with each candidate evidence and feeding them into a scoring neural network to obtain a closeness score. However, this method suffers from low efficiency due to its computational complexity of $O(m * n)$, where m and n denote the number of claims and candidate evidence, respectively. Given that the number of candidate evidence n is comparatively large, this complexity becomes impractical.

To address the aforementioned challenges, we propose a novel **R**etrieval-**A**ugmented **V**erification framework, dubbed **RAV**. RAV consists of two major modules: the hybrid evidence retrieval and the joint fact verification. The hybrid evidence retrieval involves a retriever for preliminary filtering of candidate evidence, followed by a ranker for more precise sorting of the remaining evidence. The retriever is implemented as a bi-encoder (Guo et al., 2022a), wherein the claim and evidence are processed by separate encoders. With the inference time complexity of $O(m + n)$, the retriever's efficiency can be further optimized using Approximate Nearest Neighbor techniques, making it well-suited for large-scale retrieval tasks. The ranker is implemented as a cross-encoder (Zhang et al., 2021), which offers higher precision by considering semantics from both the claim and evidence sides. In the joint fact verification module, the signals derived from the fact verification loss can be back-propagated to guide the updates of both the retriever and ranker. RAV can be easily adapted to existing fact verification models with minor modifications and boost the verification performance, demonstrating its generality and superiority.

2 Related Works

The mainstream paradigm of fact-checking is "retrieve-then-verify", and most of the existing methods focus on the claim verification phase. As for claim verification, current researches primarily adopt three distinct methodologies: sequence inference, graph reasoning, and knowledge-enhanced approaches. For sequence inference, early works concatenate all pieces of evidence for feature analysis (Nie et al., 2019). However, most current efforts adopt a nuanced strategy that integrates each evidence with the claim (Wu et al., 2021; Si et al., 2021; Kruengkrai et al., 2021), generating semantically enhanced claim-evidence pairs to facilitate deep interactions. For graph-based fact verification, the interplay among pieces of evidence is achieved via the mechanism of information dissemination across graph structures (Yang et al., 2020; Shi et al., 2021; Wang et al., 2022). Depending on the granularity of node attributes, evidence graph can be delineated into sentence-level, token-level, and multi-level. Furthermore, in recent years, some works achieve the augmentation of evidence information through the integration of text description (Yan et al., 2024; Zhao et al., 2022) or external knowledge (Zou et al., 2023; Kim et al., 2023), thereby effectively improving the precision of fact verification.

Despite the substantial advancements achieved in the phase of claim verification, the majority of existing studies fail to optimize the evidence retrieval module during the training process A pre-trained classifier is utilized to reassess evidence from optimized retriever in ReRead (Hu et al., 2023), undergoing fine-tuning through joint training (Li et al., 2017). However, this approach needs manually annotated gold evidence for supervisory information and fails to address the efficiency optimization in large-scale retrieval (Yin and Roth, 2018; Zhang et al., 2023; Li et al., 2019).

Evidence Retrieval provides a data foundation for claim verification to ensure the accuracy of verification results. Existing methods are mostly KNN-based, which can be subdivided into evidence retrieval based on feature engineering (Chakrabarty et al., 2018), neural ranking (Subramanian and Lee, 2020; Aly and Vlachos, 2022), and pre-trained models (Liu et al., 2020). These strategies depend on specific heuristic techniques, yet the evidence thus gathered may not benefit the process of claim verification well. GERE (Chen et al., 2022b) introduces a generative evidence retrieval mechanism to reduce computational resource consumption, but it remains an independent retrieval process. Hence, it can be concluded that there remains an absence

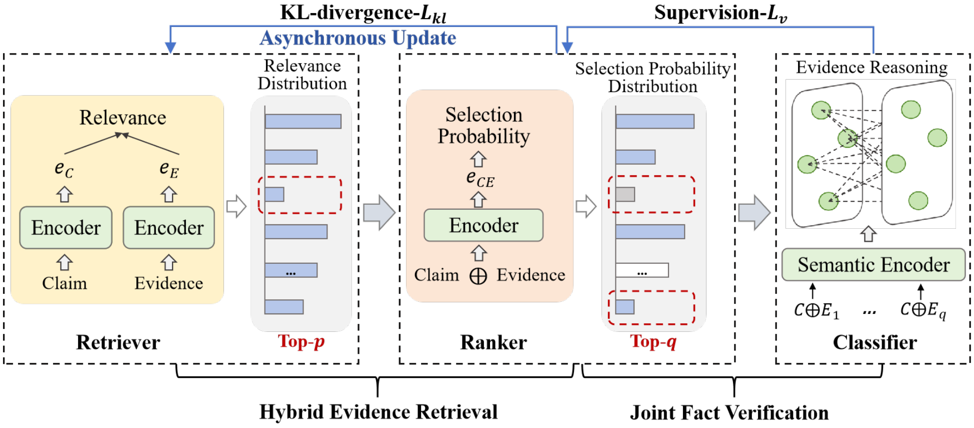


Figure 2: Framework of the proposed RAV model.

of a genuinely universal, end-to-end framework.

3 Methodology

As depicted in Figure 2, the proposed RAV comprises two phases: hybrid evidence retrieval and joint fact verification. Given a claim c and n pieces of candidate evidence $E = \{e_1, e_2, \ldots, e_n\}$, the hybrid evidence retrieval selects the top q pieces of evidence ($q \ll n$), while the fact verification stage endeavors to categorize the claim.

3.1 Hybrid Evidence Retrieval

Retriever. Due to the considerable quantity of candidate evidence, retriever is designed to efficiently filter out irrelevant candidates and recall valuable evidence. As illustrated in the left segment of Figure 2, retriever is implemented as a bi-encoder model (Zhang et al., 2021). Retriever comprises two distinct encoders: one dedicated to encoding the input claim and the other tasked with encoding the candidate evidence. The deliberate separation of encoding processes for claim and evidence yields a notable reduction in inference times, rendering it well-suited for real-time applications and large-scale endeavors.

Given the input claim c and the candidate evidence set $E = \{e_i\}$, the feed-forward process of retriever is formally defined as:

$$h_c = \text{Encoder}_c(c), \quad h_{ei} = \text{Encoder}_e(e_i),$$

in which Encoder_c and Encoder_e denote the encoder for claim and evidence, respectively. The pairwise cosine similarity is further calculated as:

$$\hat{S}_n = \{\hat{s}_1, \cdots, \hat{s}_n\} = \{\cos(h_c, h_{e1}), \cdots, \cos(h_c, h_{en})\},$$

where $\hat{s}_i$ denotes the cosine similarity between the claim c and the evidence e_i. Then, p evidence with the highest similarity $\hat{s}_i$ is selected as the output of retriever $\hat{E} = \{\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_p\}$. The relevance distribution of evidence in $\hat{E}$ is defined as: $\hat{S}_p = \{\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_p\}$.

Ranker. Following the efficient reduction of candidate evidence number from n to p by the retriever, the ranker aims to meticulously select q evidence pieces from the remaining pool with $q < p \ll n$. Ranker aims to identify and prioritize the most pertinent evidence from the subset generated by the retriever. To capture the semantic correlations between claims and evidence effectively, the ranker is implemented as a cross-encoder (Zhang et al., 2021). As illustrated in Figure 2, the ranker takes both the claim and the evidence as input, generating a joint representation. Although more computationally intensive, the cross-encoder framework enables the ranker to make informed decisions by considering the holistic context from both sides.

Based on the claim c and remained evidence $\hat{E}$, the forward process of ranker is as follows:

$$\tilde{h}_j = \text{Encoder}_r(c : \hat{e}_j), \quad \tilde{s}_j = \frac{\exp(\tilde{h}_j)}{\sum_{\hat{e}_i \in \hat{E}} \exp(\tilde{h}_i)},$$

in which ":" denotes the concatenation, and $\tilde{s}_j$ serves as the criteria for selecting the top q evidence. For evidence in $\hat{E}$, the ranker calculates its selection probability as $\tilde{S} = \{\tilde{s}_1, \tilde{s}_2, \ldots, \tilde{s}_p\}$.

Joint Optimization. The separation of the retriever and ranker renders the gradient intractable for backpropagation to the retriever. Here we utilize the KL-divergence between the relevance distribution $\hat{S}$ and the selection probability distribution $\tilde{S}$ as a supervisory signal to optimize the retriever:

$$\mathcal{L}_{kl} = D_{KL}(\tilde{S} \parallel \hat{S}) = \sum_{i=1}^{p} \hat{s}_i \ln\left(\frac{\hat{s}_i}{\tilde{s}_i}\right).$$

By leveraging KL-divergence, the discrepancy between the output distributions of the retriever and the ranker is minimized, facilitating the integration of knowledge from ranker to update retriever.

3.2 Joint Fact Verification

RAV is a universal framework capable of integrating various claim verification models. Here the popular graph-based model GEAR (Zhou et al., 2019) is taken as an example. Following the architectural principles of GEAR, a fully-connected evidence graph is constructed, where each node corresponds to an individual piece of evidence. The claim is concatenated with each retrieved evidence, forming the evidence representation. Following the graph-based message-passing and pooling operations, the combined representation of the claim and evidence is obtained, which is then fed into a classifier to derive the final decision. To utilize the verification module to provide supervisory signals for original unlabeled evidence retrieval, the ranker and classifier are jointly optimized as:

$$\mathcal{L}_v = -(y \log(y^*) + (1-y) \log(1-y^*)),$$

where $y \in \{0, 1, 2\}$ denotes the truth label of each claim, and y^* is the predicted label.

3.3 Model Training Paradigm

To improve overall efficiency, we introduce asynchronous parameter updates within the hybrid evidence retrieval framework. The training batches are partitioned into distinct iterations for optimized processing. Within each iteration, retriever parameters are fixed, and joint training between the ranker and classifier is conducted based on $\mathcal{L}_v$. Subsequently, in the final step of each iteration, parameters of both the ranker and classifier are fixed, and asynchronous optimization using $\mathcal{L}_{kl}$ is implemented for the retriever. The detailed training process is outlined in Algorithm 1 provided in Appendix A.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our proposals on the large scale dataset FEVER (Thorne et al., 2018), which consists of 185,455 annotated claims with 5,416,537 Wikipedia documents.

Evaluation Metrics. Following previous works, we use Label Accuracy (LA) and FEVER score as the evaluation metrics for claim verification (Hanselowski et al., 2018; Zhou et al., 2019; Chen et al., 2022a). Besides, giving gold evidence, Precision, Recall and F1 are used to evaluate evidence retrieval results (Chen et al., 2022c).

	Dev		Test	
Models	**LA**	**FEVER**	**LA**	**FEVER**
BERT Concat	73.67	68.89	71.01	65.64
BERT Concat+ GERE	74.41	70.25	71.83	66.40
BERT Concat+ RAV	**75.88**	**72.48**	**72.23**	**68.34**
GAT	76.13	71.04	72.03	67.56
GAT+ GERE	77.09	72.36	72.81	69.40
GAT+ RAV	**79.36**	**74.79**	**76.63**	**73.47**
GEAR	74.84	70.69	71.60	67.10
GEAR+ GERE	75.96	71.88	72.52	68.34
GEAR+ RAV	**80.89**	**74.93**	**79.91**	**74.19**
KGAT	78.29	76.11	74.07	70.38
KGAT+ GERE	79.44	77.38	75.24	71.17
KGAT+ RAV	**81.48**	**78.53**	**80.23**	**76.45**

Table 1: Performance on claim verification.

	Dev			Test		
Models	**P**	**R**	**F1**	**P**	**R**	**F1**
TF-IDF	-	-	17.20	11.28	47.87	18.26
ESIM	24.08	86.72	37.69	23.51	84.66	36.80
BERT	27.29	**94.37**	42.34	25.21	87.47	39.14
XLNet	26.60	87.33	40.79	25.55	85.34	39.33
RoBERTa	26.67	87.64	40.90	25.63	85.57	39.45
RAV	**39.47**	87.89	**54.48**	**39.20**	**87.77**	**54.20**

Table 2: Performance on evidence retrieval.

Please refer to the Appendix B for the detailed implementation details due to the limited pages.

4.2 Performance on Claim Verification

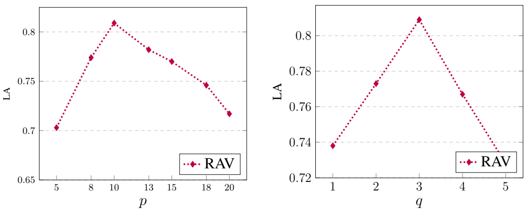
We select GERE (Chen et al., 2022c) as the primary baseline, an advanced generative evidence retrieval method. In line with GERE, we incorporate several established fact verification models as base models: BERT Concat (Zhou et al., 2019), GAT (Liu et al., 2020), KGAT (Liu et al., 2020), and GEAR (Zhou et al., 2019). As illustrated in Table 1, our framework demonstrates superior performance compared to GERE. Moreover, the application of our hybrid joint framework to existing claim verification models effectively enhances their overall performance, underscoring the superiority of our framework.

4.3 Performance on Evidence Retrieval

Following previous works (Chen et al., 2022c), we select several representative evidence retrieval methods as baselines, including TF-IDF (Thorne et al., 2018), ESIM (Hanselowski et al., 2018), BERT (Liu et al., 2020), XLNet (Zhong et al., 2020) and RoBERTa (Zhong et al., 2020). As depicted in Table 2, RAV demonstrates superior performance compared to nearly all baseline meth-

	Dev		Test	
Models	LA	FEVER	LA	FEVER
GEAR	74.84	70.69	71.60	67.10
-w/o Ranker	75.43	71.63	73.83	68.92
-w/o Retriever	79.37	74.44	77.36	72.42
GEAR + RAV	**80.89**	**74.93**	**79.91**	**74.19**

Table 3: Performance of ablation study.



(a) Accuracy vs. p (b) Accuracy vs. q

Figure 3: Hyperparameter sensitivity analysis.

ods. While KGAT exhibits a marginally higher recall rate, RAV surpasses it significantly in terms of precision and F1 score.

4.4 Ablation Study

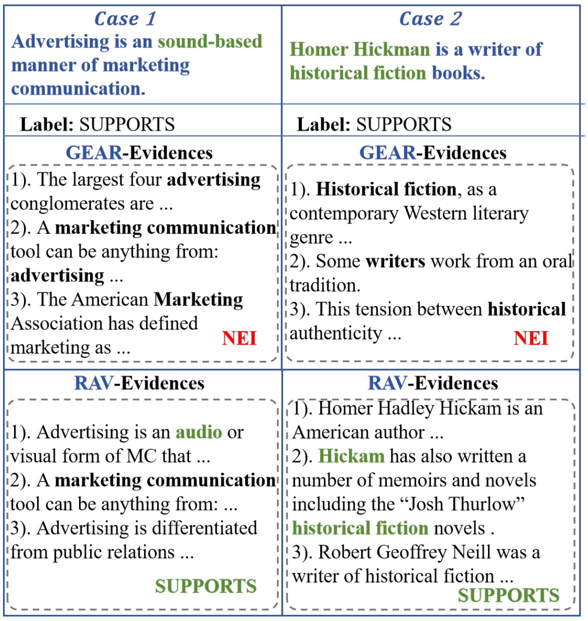
We design ablation studies to verify the effectiveness of core modules by removing the ranker and retriever, respectively. As shown in Table 3, our model outperforms both -w/o Retriever and -w/o Ranker variations, indicating that both the asynchronous updating mechanism of the retriever supervised by the ranker and the joint training framework between the ranker and classifier contribute to the efficacy of fact verification. Comparatively, -w/o Retriever achieves superior performance over -w/o Ranker, indicating that the cross-encoder is particularly advantageous for the verification task.

4.5 Hyperparameter Sensitivity Analysis

The hyperparameters p and q dictate the number of evidence retrieved by the retriever and ranker, respectively. As depicted in Figure 3, excessively large or small values lead to the performance decline. Retaining an excessive number of evidence instances may introduce unnecessary noise, while preserving too few pieces of evidence risks filtering out crucial information.

4.6 Efficiency Study

The bi-encoder excels in rapid data processing, while the cross-encoder trades off some efficiency for improved accuracy. Consequently, variation models employing the bi-encoder (-w/o Ranker) are expected to be more efficient than -w/o Retriever. As anticipated, under identical settings, the training and inference time is 92.2, 54.2 and 65.5 minutes for -w/o Retriever, -w/o Ranker, and RAV, respectively, validating this theoretical hypothesis.



Case 1 **Advertising is an sound-based manner of marketing communication.**	*Case 2* **Homer Hickman is a writer of historical fiction books.**
Label: SUPPORTS	**Label:** SUPPORTS
GEAR-Evidences	**GEAR-Evidences**
1). The largest four **advertising** conglomerates are ... 2). A **marketing communication** tool can be anything from: **advertising** ... 3). The American **Marketing** Association has defined marketing as ... **NEI**	1). **Historical fiction**, as a contemporary Western literary genre ... 2). Some **writers** work from an oral tradition. 3). This tension between **historical** authenticity ... **NEI**
RAV-Evidences	**RAV-Evidences**
1). Advertising is an **audio** or visual form of MC that ... 2). A **marketing communication** tool can be anything from: ... 3). Advertising is differentiated from public relations ... **SUPPORTS**	1). Homer Hadley Hickam is an American author ... 2). **Hickam** has also written a number of memoirs and novels including the “Josh Thurlow” **historical fiction** novels . 3). Robert Geoffrey Neill was a writer of historical fiction ... **SUPPORTS**

Figure 4: Cases Studies.

4.7 Case Study

Compared to KNN-based method, our proposed evidence retrieval approach can extract implicit evidence. As shown in Figure 4, the expressions "*sound-based*" and "*audio*" are synonymous. KNN-based method struggles to capture such features, whereas our approach benefits from a joint training process, enabling it to explicitly assign a higher selection probability to the evidence that are more favorable for claim verification. Similarly, in *case 2*, the key evidence is inferred through the statements "$Hickam\ has\ written\ Josh\ Thurlow.$" and "$Josh\ Thurlow\ is\ a\ historical\ fiction\ novel.$". KNN-based method also fails to capture these nuanced evidence.

5 Conclusion

In this paper, we delves into the critical issue of evidence retrieval and puts forth a joint fact verification framework featuring hybrid evidence retrieval. Experimental results show that integrating RAV with established claim verification models markedly boosts overall performance.

Limitations

Initially, there is potential for our model to realize improvements in efficiency. The utilization of bi-encoders coupled with approximated nearest neighbor searching (ANN) techniques has the capacity to enhance retrieval efficiency for a broader candidate set. Additionally, in this study, we refrained from making further enhancements to the claim verification module, choosing instead to integrate the existing models into our framework. The incorporation of a more sophisticated evidence reasoning approach holds the promise of generating enhanced supervisory signals for the evidence retrieval process, thus potentially elevating the model's overall efficacy to a certain extent.

Acknowledgements

This work was partially sponsored by CAAI-Huawei MindSpore Open Fund, Compute services from Hebei Artificial Intelligence Computing Center. And it was partly supported by the Natural Science Foundation of China (No.62372057), the Fundamental Research Funds for the Central Universities (No.500422828), the Beijing Natural Science Foundation (No.4244083).

References

Rami Aly and Andreas Vlachos. 2022. Natural logic-guided autoregressive multi-hop document retrieval for fact verification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6123–6135, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tuhin Chakrabarty, Tariq Alhindi, and Smaranda Muresan. 2018. Robust document retrieval and individual evidence modeling for fact extraction and verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 127–131, Brussels, Belgium. Association for Computational Linguistics.

Jiangjie Chen, Qiaoben Bao, Changzhi Sun, Xinbo Zhang, Jiaze Chen, Hao Zhou, Yanghua Xiao, and Lei Li. 2022a. Loren: Logic-regularized reasoning for interpretable fact verification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10482–10491.

Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2022b. Gere: Generative evidence retrieval for fact verification. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2184–2189.

Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2022c. Gere: Generative evidence retrieval for fact verification. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2184–2189, New York, NY, USA. Association for Computing Machinery.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

Jiafeng Guo, Yinqiong Cai, Yixing Fan, Fei Sun, Ruqing Zhang, and Xueqi Cheng. 2022a. Semantic models for the first-stage retrieval: A comprehensive review. *ACM Trans. Inf. Syst.*, 40(4).

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022b. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. UKP-athene: Multi-sentence textual entailment for claim verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108, Brussels, Belgium. Association for Computational Linguistics.

Xuming Hu, Zhaochen Hong, Zhijiang Guo, Lijie Wen, and Philip Yu. 2023. Read it twice: Towards faithfully interpretable fact verification by revisiting evidence. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2319–2323, New York, NY, USA. Association for Computing Machinery.

Jiho Kim, Sungjin Park, Yeonsu Kwon, Yohan Jo, James Thorne, and Edward Choi. 2023. FactKG: Fact verification via reasoning on knowledge graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16190–16206, Toronto, Canada. Association for Computational Linguistics.

Canasai Kruengkrai, Junichi Yamagishi, and Xin Wang. 2021. A multi-level attention model for evidence-based fact checking. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2447–2460, Online. Association for Computational Linguistics.

Chaozhuo Li, Senzhang Wang, Yukun Wang, Philip Yu, Yanbo Liang, Yun Liu, and Zhoujun Li. 2019. Adversarial learning for weakly-supervised social network alignment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 996–1003.

Chaozhuo Li, Senzhang Wang, Dejian Yang, Zhoujun Li, Yang Yang, Xiaoming Zhang, and Jianshe Zhou.

2017. Ppne: property preserving network embedding. In *Database Systems for Advanced Applications: 22nd International Conference, DASFAA 2017, Suzhou, China, March 27-30, 2017, Proceedings, Part I 22*, pages 163–179. Springer.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351, Online. Association for Computational Linguistics.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:6859–6866.

Qi Shi, Yu Zhang, Qingyu Yin, and Ting Liu. 2021. Logic-level evidence retrieval and graph-based verification network for table-based fact verification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiasheng Si, Deyu Zhou, Tongzhe Li, Xingyu Shi, and Yulan He. 2021. Topic-aware evidence reasoning and stance-aware aggregation for fact verification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1612–1622, Online. Association for Computational Linguistics.

Shyam Subramanian and Kyumin Lee. 2020. Hierarchical Evidence Set Modeling for automated fact extraction and verification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7798–7809, Online. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. 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 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Yiqi Wang, Chaozhuo Li, Zheng Liu, Mingzheng Li, Jiliang Tang, Xing Xie, Lei Chen, and Philip S Yu. 2022. An adaptive graph pre-training framework for localized collaborative filtering. *ACM Transactions on Information Systems*, 41(2):1–27.

Lianwei Wu, Rao Yuan, Ling Sun, and Wangbo He. 2021. Evidence inference networks for interpretable claim verification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:14058–14066.

Hao Yan, Chaozhuo Li, Ruosong Long, Chao Yan, Jianan Zhao, Wenwen Zhuang, Jun Yin, Peiyan Zhang, Weihao Han, Hao Sun, et al. 2024. A comprehensive study on text-attributed graphs: Benchmarking and rethinking. *Advances in Neural Information Processing Systems*, 36.

Xiaoyu Yang, Feng Nie, Yufei Feng, Quan Liu, Zhigang Chen, and Xiaodan Zhu. 2020. Program enhanced fact verification with verbalization and graph attention network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7810–7825, Online. Association for Computational Linguistics.

Wenpeng Yin and Dan Roth. 2018. TwoWingOS: A two-wing optimization strategy for evidential claim verification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 105–114, Brussels, Belgium. Association for Computational Linguistics.

Hang Zhang, Yeyun Gong, Yelong Shen, Jiancheng Lv, Nan Duan, and Weizhu Chen. 2021. Adversarial retriever-ranker for dense text retrieval. *ArXiv*, abs/2110.03611.

Hengran Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2023. From relevance to utility: Evidence retrieval with feedback for fact verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6373–6384, Singapore. Association for Computational Linguistics.

Jianan Zhao, Meng Qu, Chaozhuo Li, Hao Yan, Qian Liu, Rui Li, Xing Xie, and Jian Tang. 2022. Learning on large-scale text-attributed graphs via variational inference. *arXiv preprint arXiv:2210.14709*.

Yi Zhao, Chaozhuo Li, Jiquan Peng, Xiaohan Fang, Feiran Huang, Senzhang Wang, Xing Xie, and Jibing Gong. 2023. Beyond the overlapping users: Cross-domain recommendation via adaptive anchor link learning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1488–1497.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Reasoning over semantic-level graph for fact checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6170–6180, Online. Association for Computational Linguistics.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. GEAR: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901, Florence, Italy. Association for Computational Linguistics.

Anni Zou, Zhuosheng Zhang, and Hai Zhao. 2023. Decker: Double check with heterogeneous knowledge for commonsense fact verification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11891–11904, Toronto, Canada. Association for Computational Linguistics.

A Algorithm

Algorithm 1 Training Process of RAV

```
Input: Claim $c$ and candidate evidence $E = \{e_1, e_2, \dots, e_n\}$.
Output: Predicted label $y^*$.
$steps \leftarrow t$
for i in range(t) do
    if $i \leq t - 1$ then
        {Ranker.train() and Retriever.eval()}
        $\hat{E} \leftarrow \{\hat{e}_1, \hat{e}_2, \dots, \hat{e}_p\}$
        $\tilde{E} \leftarrow \{\tilde{e}_1, \tilde{e}_2, \dots, \tilde{e}_q\}$
        $y^* \leftarrow Ranker(c, \tilde{E})$
        $L_{label} \leftarrow CrossEntropy(y, y^*)$
        Ranker parameters update.
    end if
    if $i == t - 1$ then
        {Ranker.train() and Retriever.eval()}
        $\hat{S} \leftarrow \{\hat{s}_1, \hat{s}_2, \dots, \hat{s}_p\}$
        $\tilde{S} \leftarrow \{\tilde{s}_1, \tilde{s}_2, \dots, \tilde{s}_p\}$
        $L_{kl} \leftarrow D_{KL}(\hat{S} \parallel \tilde{S})$
        Ranker parameters update.
    end if
end for
Return: $y^*$
```

Algorithm 1 describes the training process of RAV in one iteration. For simplification, *Ranker* represents the joint model of ranker and classifier, and there are t training steps in each iteration. At each step, *Ranker* will be jointly optimized with the retriever parameters fixed. $\hat{E}$ and $\tilde{E}$ represent the evidence set retrieved by retriever and ranker, respectively. Afterwards, with the claim c and evidence $\tilde{E}$ as input, the classifier generates the predicted label y^* through evidence reasoning. The cross entropy loss serves as the optimization objective of both ranker and classifier by back-propagating gradients to the two stages. In addition, to enhance the overall efficiency, the retriever will be asynchronous updated only at the last step. The KL-divergence between the relevance distribution $\hat{S}$ and the selection probability distribution $\tilde{S}$ is utilized as the supervisory signal to optimize the retriever.

B Implementation details

In the data preprocessing stage, we adopt the *entity linking* approach (Hanselowski et al., 2018) to select 20 related documents for each claim. Furthermore, we use the *modified ESIM model* (Chen et al., 2017) to generate 30 candidate evidence from the selected documents.