Read it Twice: Towards Faithfully Interpretable Fact Verification by Revisiting Evidence

Xuming Hu
Tsinghua University
hxm19@mails.tsinghua.edu.cn

Zhaochen Hong
Tsinghua University
hongzc20@mails.tsinghua.edu.cn

Zhijiang Guo
University of Cambridge
zg283@cam.ac.uk

Lijie Wen
Tsinghua University
wenlj@tsinghua.edu.cn

Philip S. Yu
University of Illinois at Chicago
psyu@cs.uic.edu

ABSTRACT

Real-world fact verification task aims to verify the factuality of a claim by retrieving evidence from the source document. The quality of the retrieved evidence plays an important role in claim verification. Ideally, the retrieved evidence should be *faithful* (reflecting the model's decision-making process in claim verification) and *plausible* (convincing to humans), and can improve the *accuracy* of verification task. Although existing approaches leverage the similarity measure of semantic or surface form between claims and documents to retrieve evidence, they all rely on certain heuristics that prevent them from satisfying all three requirements. In light of this, we propose a fact verification model named ReRead to retrieve evidence and verify claim that: (1) Train the evidence retriever to obtain interpretable evidence (i.e., faithfulness and plausibility criteria); (2) Train the claim verifier to revisit the evidence retrieved by the optimized evidence retriever to improve the accuracy. The proposed system is able to achieve significant improvements upon best-reported models under different settings.

CCS CONCEPTS

• **Information systems → Information systems applications**.

KEYWORDS

Automated Fact-Checking, Real-world Systems, Latent Variable Models, Evidence Retrieval, Claim Verification

ACM Reference Format:
Xuming Hu, Zhaochen Hong, Zhijiang Guo, Lijie Wen, and Philip S. 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), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3592049

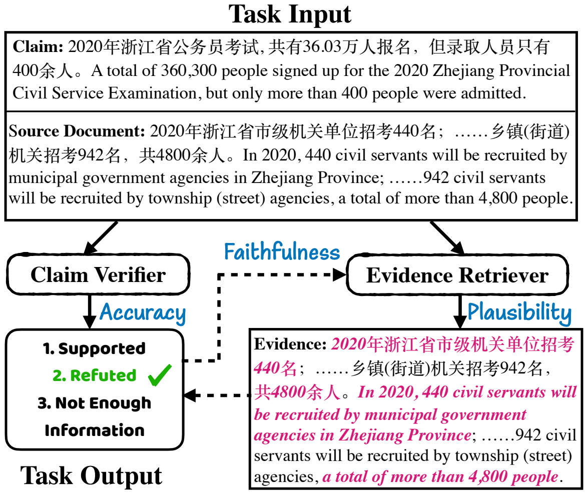
Task Input
Claim: 2020年浙江省公务员考试, 共有36.03万人报名，但录取人员只有400余人。A total of 360,300 people signed up for the 2020 Zhejiang Provincial Civil Service Examination, but only more than 400 people were admitted.
Source Document: 2020年浙江省市级机关单位招考440名；……乡镇(街道)机关招考942名，共4800余人。In 2020, 440 civil servants will be recruited by municipal government agencies in Zhejiang Province; ……942 civil servants will be recruited by township (street) agencies, a total of more than 4,800 people.
Claim Verifier
Faithfulness
Evidence Retriever
Accuracy
Plausibility
1. Supported
2. Refuted ✓
3. Not Enough Information
Evidence: 2020年浙江省市级机关单位招考440名；……乡镇(街道)机关招考942名，共4800余人。In 2020, 440 civil servants will be recruited by municipal government agencies in Zhejiang Province; ……942 civil servants will be recruited by township (street) agencies, a total of more than 4,800 people.
Task Output

Figure 1: A case of ReRead. The evidence retriever should retrieved evidence which could give the plausible reason why the verification result is "Refuted" and reflect the verifier's decision-making process. With the training of the evidence retriever, it can provide the verifier with better evidence to revisit and improve the accuracy of the fact verification task.

1 INTRODUCTION

The spread of misinformation has become a significant issue in today's society, particularly in the digital age where information can be easily disseminated and shared across various platforms [3, 24, 33]. As such, fact verification has emerged as a crucial task in combating this issue by assessing the factuality of claims made in written or spoken language [1, 4, 7, 22, 32]. To achieve this goal, it is essential to have appropriate evidence that supports or refutes a claim. Therefore, how to retrieve suitable evidence from a large number of source documents is a key component of fact verification.

As shown in Figure 1, a real-world claim from Chinese social media and corresponding source document are retrieved through Google search engine. We need to retrieve *faithful* (reflecting the decision-making process of the verifier in claim verification) and *plausible* (explaining the reason for the factuality of the claim) evidence from the noisy document to improve the task *accuracy* of claim verification [8, 36]. In this case, evidence such as "more than 4800 people" needs to be retrieved to counter the claim of "only more than 400 people". Although evidence plays a crucial role in fact verification, early automated fact verification attempts disregarded this, and solely relied on the surface patterns of the claim to verify it while ignoring the information that evidence provides [25, 31]. Consequently, these approaches were unable to identify well-camouflaged misinformation [26]. Recent efforts to address this issue involve asking annotators to create claims and evidence by mutating sentences from Wikipedia articles [2, 28]. However,

SIGIR '23July 23–27, 2023Taipei, Taiwan.
2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3592049

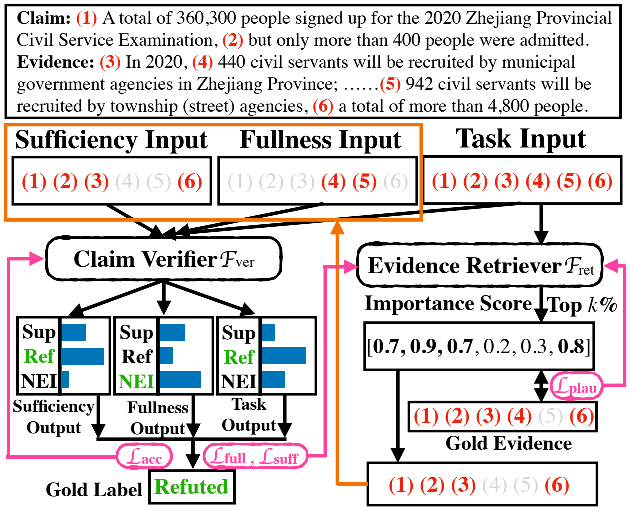
Claim: (1) A total of 360,300 people signed up for the 2020 Zhejiang Provincial Civil Service Examination, (2) but only more than 400 people were admitted.
Evidence: (3) In 2020, (4) 440 civil servants will be recruited by municipal government agencies in Zhejiang Province; ……(5) 942 civil servants will be recruited by township (street) agencies, (6) a total of more than 4,800 people.
Sufficiency Input
Fullness Input
Task Input
(1) (2) (3) (4) (5) (6)
(1) (2) (3) (4) (5) (6)
(1) (2) (3) (4) (5) (6)
Claim Verifier $\mathcal{F}_{ver}$
Evidence Retriever $\mathcal{F}_{ret}$
Sup
Ref
NEI
Sufficiency Output
Fullness Output
Task Output
$\mathcal{L}_{acc}$
$\mathcal{L}_{full}$, $\mathcal{L}_{suff}$
Gold Label
Refuted
Importance Score
Top $k\%$
[0.7, 0.9, 0.7, 0.2, 0.3, 0.8]
$\mathcal{L}_{plau}$
(1) (2) (3) (4) (5) (6)
Gold Evidence
(1) (2) (3) (4) (5) (6)

Figure 2: Architecture of ReRead.

these synthetic claims generated from Wikipedia cannot serve as a substitute for real-world claims that circulate in the media ecosystem. As a result, other works resorted to scraping claims from fact-checking sites and using search engines to find supporting documents [9, 10, 34]. However, the source documents retrieved in this way is often noisy, which hinders the accuracy of verification task. To address this, Hu et al. [10] retrieve relevant evidence from the source documents by measuring semantic similarity between the claim and the evidence and Gupta and Srikumar [9] develop an attention-based evidence aggregation model. However, these methods all rely on certain heuristics and cannot simultaneously satisfy the three requirements of being faithful, plausible, and improving the fact verification accuracy.

We propose the novel real-world fact verification model ReRead, which meets three key requirements by: (1) Training an evidence retriever for interpretable evidence based on faithfulness and plausibility criteria; (2) Training a claim verifier to re-evaluate evidence from the optimized retriever, enhancing accuracy. As illustrated in Figure 1, ReRead fine-tunes the verifier using labeled data, then utilizes it to help the retriever obtain faithful evidence. The retriever also uses gold evidence to boost plausibility. Improved evidence provided by the trained retriever allows the verifier to refine accuracy. Our main contributions include: (1) A novel model for retrieving faithful and plausible evidence, increasing verification accuracy; (2) Experiments demonstrating a 4.31% F1 performance gain over the SOTA baseline on a real-world dataset, with extensive analysis validating ReRead's effectiveness.

2 TRAINING GOAL ANALYSIS

We have three training goals: (1) The retrieved evidence needs to have **Faithfulness**, which means how accurately the evidence reflects the true reasoning process of the verifier to predict the verification label [14]. We use two metrics: **Fullness** reflects the change in probability of the predicted label after removing evidence from the source document. **Sufficiency** reflects the probability change of using only evidence to predict the label, in other words, if the evidence is really influential, the probability of the label will not change significantly. (2) The retrieved evidence needs to have **Plausibility** to convince the verifier's prediction [6]. We adopt gold evidence to train the retrieved evidence. (3) The **Accuracy** of the task needs to be improved by revisiting the evidence retrieved.

3 MODEL ARCHITECTURE

As shown in Figure 2, ReRead first leverage the labeled data to fine tune the claim verifier with $\mathcal{L}_{acc}$.ReRead utilizes gold evidence to provide plausibility of the retrieved evidence ($\mathcal{L}_{plau}$) and gold labels to provide faithfulness of evidence ($\mathcal{L}_{full}$ and $\mathcal{L}_{suff}$).

3.1 Sentence Encoder

We adopt the BERT encoder [5] to obtain the semantic embeddings of each sentence within the claim and source document. For a given claim C and its corresponding source document D, we get their sentence embeddings by adding a special token [CLS] at the beginning of each sentence and utilizing the [CLS] position embeddings. This produces an embedding matrix $S_{emb} \in \mathbb{R}^{l \times d}$ for the claim and document, where l is number of total sentences and $d = 768$.

3.2 Claim Verifier

Our claim verifier takes S_{emb} as input and classifies the claim into three categories: refuted (Ref), supported (Sup) and not enough information (NEI). During training, the verifier performs classification based on the claim and the document.

We use a neural network-based classifier $\mathcal{F}_{ver}$ to achieve this. It takes S_{emb} as input and outputs a probability prediction vector $\mathcal{F}_{ver}(S_{emb}) = (p_{Sup}, p_{Ref}, p_{NEI})^\top$, where p_{Sup}, p_{Ref} and p_{NEI} represent the probability of claim Sup, Ref, or NEI, respectively. We denote the verification result as random variable v.

3.2.1 Accuracy. We adopt the criterion of accuracy to train the claim verifier to perform claim verification. To evaluate its performance, we use cross entropy loss $\mathcal{L}_{CE}(\mathcal{F}_{ver}(S_{emb}), y^*)$, which calculates the difference between the verifier's probability prediction $\mathcal{F}_{ver}(S_{emb})$ and the ground truth label $y^* \in \{0, 1, 2\}$ which indicates the Ref, Sup, and NEI, respectively. Consequently, we define the accuracy loss function as:

$$\mathcal{L}_{acc} = \mathcal{L}_{CE}(\mathcal{F}_{ver}(S_{emb}), y^*), \tag{1}$$

which is used to train the claim verifier and the sentence encoder.

3.3 Evidence Retriever

After the claim verifier is trained, the evidence retriever will be trained to improve the faithfulness of the retrieved evidence using the trained verifier and ensure plausibility using the gold evidence in the dataset. The optimized evidence further enhances the performance of verification. To achieve this, we use a neural network-based classifier $\mathcal{F}_{ret}$ and the output of the sentence encoder to obtain semantic information. Notationally, $\mathcal{F}_{ret}$ takes S_{emb} as input from the sentence encoder and outputs a vector $\mathcal{F}_{ret}(S_{emb}) \in [0, 1]^l$, which quantifies the probability that each of the l sentences in the document is important to claim verification. We denote $1, 0$ to indicate sentences are selected or not, respectively. We denote the sentence embedding obtained after passing the selected evidence to the sentence encoder as E_{emb}.

To ensure faithfulness, we use the criteria of fullness and sufficiency. For more plausible evidence, we employ the criterion of plausibility, which incentivizes the retriever to have a evidence selection that makes sense to humans. We denote the loss function for fullness, sufficiency, plausibility as $\mathcal{L}_{full}$, $\mathcal{L}_{suff}$, and $\mathcal{L}_{plau}$ respectively. Consequently, we can use $\mathcal{L}$ to jointly represent the three loss functions as the target function for the evidence retriever:

$$\mathcal{L} = \alpha_{full}\mathcal{L}_{full} + \alpha_{suff}\mathcal{L}_{suff} + \alpha_{plau}\mathcal{L}_{plau}. \tag{2}$$

3.3.1 Plausibility. We introduce the plausibility criterion to measure and enhance the degree to which evidence is plausible to humans. To select the sentences that are most important to the claim verifier, we use a Top k algorithm that selects the sentences with the highest probability scores. Specifically, we select the Top $k\%$ sentences in the document based on their probability scores. The selected evidence is denoted as E.

We adopt the claim with corresponding gold evidence and measure the difference between the predicted evidence and the gold evidence with binary cross entropy loss. We denote $\boldsymbol{g_i} \in \{0, 1\}^{|S|}$ as the gold evidence, where 0 or 1 represents whether a sentence is selected or not. The plausibility loss function could be defined as:

$$\mathcal{L}_{plau} = \mathcal{L}_{BCE}(\mathcal{F}_{ret}(S_{emb}), \boldsymbol{g_i}), \tag{3}$$

which could encourage the retriever to select evidence sentences that are more plausible during training.

3.3.2 Faithfulness-Fullness. If removing certain sentences from the document would lead to incorrect verification result, we can assume that these sentences contain critical evidence that plays a crucial role in the verification outcome. To choose the most crucial evidence, we should identify the sentences that, if removed, would significantly reduce the claim verifier's performance.

We use cross entropy loss $\mathcal{L}_{CE}(\mathcal{F}_{ver}(S_{emb}), y^*)$ to measure the verification performance, where the label y^* indicates one of three categories. To assess the impact of removing evidence sentences, we can compare the performance of $S_{emb}\backslash E_{emb}$ to the original input. Specifically, we can measure the influence of removing evidence sentences with the following formula:

$$\mathcal{L}_{full} = \mathcal{L}_{CE}(\mathcal{F}_{ver}(S_{emb}), y^*) - \mathcal{L}_{CE}(\mathcal{F}_{ver}(S_{emb}\backslash E_{emb}, y^*). \tag{4}$$

The loss function $\mathcal{L}_{full}$ can encourage the retriever to select all sentences important to claim verification.

Ideally, the evidence retriever selects the key evidence sentences that play an decisive part in the verification process so that $\mathcal{L}_{full} < 0$. To address this issue, we can first set $\mathcal{L'}_{full}$ to 0 when corresponding $\mathcal{L}_{full} < -B_f$, where $B_f > 0$ is a hyperparameter. To transform the range of the original loss values so that it is always 0 or more, we can denote $\mathcal{L'}_{full} = \mathcal{L}_{full} + B_f$ when $\mathcal{L}_{full} > -B_f$ so that the reformulated loss value $\mathcal{L'}_{full} \geq 0$. Formally, we can define $\mathcal{L'}_{full}$ as follows:

$$\mathcal{L'}_{full} = \max(0, \mathcal{L}_{full} + B_f), \tag{5}$$

which could regulate the value of $\mathcal{L}_{full}$ into the range of $[0, +\infty)$.

3.3.3 Faithfulness-Sufficiency. To ensure that the selected evidence improves verification performance beyond what the original source document provides, we use the sufficiency criterion. This criterion incentivizes the retriever to select evidence that results in the greatest improvement in claim verification performance compared with using the original document alone.

More specifically, we adopt $\mathcal{L}_{CE}(\mathcal{F}_{ver}(E_{emb}), y^*)$ which represents the performance of using the evidence to replace the document, while $\mathcal{L}_{CE}(\mathcal{F}_{ver}(S_{emb}), y^*)$ stands for the original performance using the claim and the document as input to the claim verifier. Thus, we define the sufficiency loss function:

$$\mathcal{L}_{suff} = \mathcal{L}_{CE}(\mathcal{F}_{ver}(E_{emb}), y^*) - \mathcal{L}_{CE}(\mathcal{F}_{ver}(S_{emb}), y^*), \tag{6}$$

which encourages the retriever to select all important sentences that are used in the claim verification process. The loss function $\mathcal{L}_{suff}$ also have the potential to be negative when the retriever is well-trained. To avoid a negative loss function, we can employ similar measurements by setting a hyperparameter $B_s > 0$, which is large enough and transforming the range of value into $[0, +\infty)$. Therefore, we can define the sufficiency loss function as:

$$\mathcal{L'}_{suff} = \max(0, \mathcal{L}_{suff} + B_s) \tag{7}$$

The optimized retriever will retrieve better evidence, which improves the results of the verifier in Section 3.2 by revisiting it.

4 EXPERIMENTS AND ANALYSES

4.1 Setup and Baselines

Setup: Note that only CHEF [10] has marked the gold evidence for real-world claims. Although FEVER, [29], FEVER 2.0 [30], and FEVEROUS [2] annotate evidence retrieved from Wikipedia, they do not serve claims from the real-world. Therefore, we only use CHEF. To measure the effect of ReRead, we adjust the parameters on the train set, and report the results on dev and test sets of CHEF. The train/dev/test sets of CHEF have 8,002/999/999 samples respectively. CHEF also provides the google snippets as the evidence, which is the summary of the content of source documents provided by Google [9]. Following prior efforts [9, 10, 21], we adopt Micro F1 and Macro F1 as the evaluation metric. For base encoder, we adopt BERT-Base-Chinese [5] and RoBERTa-Base-Chinese [20]. We set k as 5% of all sentences in the source documents. We use BertAdam [15] with 4e-5 learning rate, warmup with 0.07 to optimize the cross entropy loss and set the batch size as 16. For simplicity, we set α_{full}, α_{suff}, and α_{plau} to 1 respectively.

Baselines: Following previous works [9, 10], we adopt two types of baselines: Pipeline and Joint systems. Pipeline systems first retrieve evidence from the documents according to the claim, and use the retrieved evidence to verify the claim. The evidence retriever and claim verification are two independent steps. We adopt (1) Google Snippets [9]. (2) Surface Ranker [2]. (3) Semantic Ranker [21]. (4) Hybrid Ranker [27]. Joint systems treat evidence extraction as a latent variable, and jointly optimize the evidence extraction process by claim verification loss. We adopt (5) Reinforcement-based Method [16]. (6) Multi-task based Method [35]. (7) Latent based Method [10]. In addition, we give (8) No evidence and (9) Gold evidence, to show lower and upper bounds for results.

4.2 Results and Analysis

Overall Performance. Table 1 shows the mean and standard deviation results with 5 runs of training and testing on dev and test sets of CHEF. We observe that the use of real-world evidence can improve the effect of claim verification, and source documents can bring more improvement than google snippets, which is related to the fact that source documents contains more information.

Table 1: Micro and Macro F1 Results of ReRead and baseline models across Test and Dev sets on CHEF.

System / Evidence			Test Set				Dev Set			
			BERT-Based Model		RoBERTa-Based Model		BERT-Based Model		RoBERTa-Based Model	
			Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1
Pipeline	No Evidence		54.46±2.89	52.49±2.44	55.34±2.68	53.22±2.59	54.76±2.35	52.97±2.12	55.73±2.06	53.61±2.17
	Google Snippets [9]		62.07±2.55	60.61±2.96	62.53±2.13	61.55±2.69	62.31±1.97	60.87±2.07	62.96±2.17	61.93±2.42
	Surface Ranker [2]		63.17±1.67	61.47±2.02	64.21±1.94	62.05±2.17	63.53±1.78	61.78±1.95	64.66±1.86	62.49±2.08
	Semantic Ranker [21]		63.47±1.71	61.94±1.66	64.35±1.76	62.24±1.52	63.73±1.68	62.42±1.49	64.71±1.45	62.59±1.38
	Hybrid Ranker [27]		63.29±1.65	61.80±2.31	63.98±1.53	61.78±1.48	63.12±1.72	61.53±1.59	64.32±1.83	62.11±1.43
Joint	Reinforce [16]	Google Snippets	63.76±1.52	61.74±1.88	64.46±1.82	62.42±1.67	63.54±1.38	61.48±1.63	64.81±1.69	62.80±1.72
		Source Documents	64.37±1.65	62.46±1.72	65.04±1.59	63.05±1.47	64.68±1.62	62.63±1.49	65.48±1.68	63.41±1.39
	Multi-task [35]	Google Snippets	62.78±1.41	61.98±2.59	64.19±1.98	62.62±1.76	62.94±1.86	62.37±1.65	64.51±1.79	63.05±1.76
		Source Documents	65.02±1.46	63.12±1.78	65.87±1.68	63.79±1.84	65.41±1.80	63.38±1.62	66.19±1.63	64.12±1.55
	Latent [10]	Google Snippets	64.45±1.68	62.52±2.23	65.11±1.86	63.14±1.82	64.71±1.69	62.80±1.48	65.08±1.62	63.50±1.77
		Source Documents	66.77±1.43	64.65±1.74	66.95±1.68	65.13±1.57	66.96±1.45	64.92±1.50	67.33±1.26	65.57±1.39
Pipeline	**ReRead**	Source Documents	**70.87±1.05**	**68.78±1.21**	**71.24±1.11**	**69.52±0.96**	**71.31±1.08**	**69.25±1.18**	**71.79±1.26**	**69.98±1.09**
	w/o $\mathcal{L}_{plau}$	Source Documents	67.67±1.32	65.84±1.46	68.03±1.35	66.11±1.48	67.96±1.57	66.04±1.51	68.14±1.42	66.31±1.56
	w/o $\mathcal{L}_{full}\&\mathcal{L}_{suff}$	Source Documents	68.24±1.42	66.15±1.39	68.58±1.50	66.39±1.44	68.53±1.32	66.31±1.53	68.70±1.44	66.59±1.37
Pipeline	Gold Evidence		**78.99±0.82**	**77.62±1.02**	**79.14±0.93**	**78.59±1.02**	**79.26±0.94**	**78.04±1.10**	**79.98±0.89**	**78.81±1.01**

Table 2: Quality of Retrieved Evidence Analysis.

Methods	Test Set				Dev Set			
	BERT-Base		RoBERTa-Base		BERT-Base		RoBERTa-Base	
	BLEU	F1	BLEU	F1	BLEU	F1	BLEU	F1
Surface	0.43	85.3	0.46	86.6	0.42	84.6	0.44	85.5
Semantic	0.53	88.1	0.55	89.5	0.52	88.4	0.56	89.4
Hybrid	0.48	87.7	0.50	88.9	0.46	87.5	0.48	88.6
Reinforce	0.63	89.6	0.66	90.4	0.62	89.3	0.64	90.3
Multi-task	0.66	90.4	0.67	91.5	0.64	90.3	0.65	90.8
Latent	0.68	90.8	0.69	91.4	0.67	90.5	0.69	91.2
ReRead	**0.84**	**95.3**	**0.86**	**95.4**	**0.85**	**95.1**	**0.87**	**95.7**

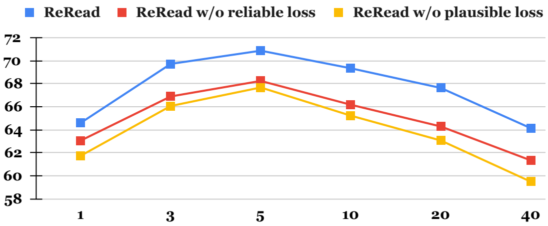

Figure 3: Micro F1 results with different k on test set.

Correspondingly, these source documents also contain more noise content, but ReRead still consistently outperforms the baselines. More specifically, compared with the previous SOTA model: Latent [10], ReRead on average achieves 4.30% higher Micro F1 and 4.32% higher Macro F1 across dev and test sets. We attribute the consistent improvement of ReRead to the faithful and plausible evidence which ReRead retrieved from source documents. ReRead is more robust than all baselines when considering standard deviations, since the evidence retriever is supervised by gold evidence through plausibility, providing higher quality evidence.

Ablation Study. We conduct an ablation study to show the effectiveness of different losses of ReRead on the dev and test sets. ReRead *w/o* $\mathcal{L}_{plau}$ means that the plausible loss function is removed, which makes the evidence retriever no longer use the gold evidence to train the selected evidence. ReRead *w/o* $\mathcal{L}_{full}\&\mathcal{L}_{suff}$ removes the faithful loss function from the claim verifier, which will cause the evidence obtained by the evidence retriever to no longer depend on the claim verification result. A general conclusion from ablation rows in Table 1 is that all losses contribute positively to the improved performance. More specifically, without $\mathcal{L}_{plau}$, the selected evidence will become unconvincing, resulting in a 3.33% F1 performance decrease. Removing the $\mathcal{L}_{full}\&\mathcal{L}_{suff}$ will select task-agnostic evidence, resulting in a 2.90% F1 performance loss.

Quality of Retrieved Evidence Analysis. We assess the retrieved evidence quality by comparing it to gold evidence in dev and test sets. We use the BLEU [23] to gauge the similarity between retrieved and gold evidence, with higher BLEU indicating better quality. Additionally, 5 Ph.D. students annotate verification labels for 100 claims based on retrieved evidence, while 2 Ph.D. students validate the data. This helps us evaluate the **interpretability** of retrieved evidence. Table 2 displays the BLEU and Micro F1 scores. ReRead shows a notable 17% BLEU improvement over the SOTA baseline, proving that incorporating plausible loss for evidence retriever training helps ReRead obtain higher-quality evidence, resulting in a 5.87% increase in human-labeled F1 verification accuracy.

Effect of the Selection Ratio k. As shown in Figure 3, we report Micro F1 scores of BERT-Base encoder against different k on the test set. A low k value may have a detrimental effect on the information sufficiency of the retrieved evidence, thus affecting the verification results. The F1 score of ReRead does not increase monotonically, as irrelevant evidence are included. The model achieves the best performance when $k = 5$, which means 5% sentences are selected as evidence is the most appropriate. If we remove the faithful and plausible loss, the F1 performance of ReRead will drop 3.24% F1 on average due to missing guidance from the gold label and evidence.

5 CONCLUSION

In this paper, we propose a novel fact verification framework ReRead, which adopt the plausibility, fullness, and sufficiency criteria to retrieve appropriate evidence from real-world documents. The retrieved evidence could reflect the factuality of the claim and convince to human. With the training of the evidence retriever, it can further provide the claim verifier with better evidence to revisit and improve the accuracy of the verification task. Experiments on real-world dataset shows the effectiveness of ReRead. In the future, we can extend the research on faithful interpretation to the construction of knowledge graphs [11–13, 19, 37], the extraction and answering of structured knowledge [17, 18].

ACKNOWLEDGMENTS

Lijie Wen is the corresponding author. Xuming Hu and Zhaochen Hong have made equal contributions to this work. Our code is available at: https://github.com/THU-BPM/ReRead. The work was supported by the National Key Research and Development Program of China (No. 2019YFB1704003), the National Nature Science Foundation of China (No. 62021002), NSF under grants III1909323, Tsinghua BNRist and Beijing Key Laboratory of Industrial Bigdata System and Application.

REFERENCES

[1] Bill Adair, Chengkai Li, Jun Yang, and Cong Yu. 2017. Progress toward "the holy grail": The continued quest to automate fact-checking. In *Proceedings of the 2017 Computation+Journalism Symposium.*

[2] Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. FEVEROUS: Fact Extraction and VERification Over Unstructured and Structured information. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1).*

[3] Bjarte Botnevik, Eirik Sakariassen, and Vinay Setty. 2020. Brenda: Browser extension for fake news detection. In *Proc. of SIGIR.* 2117–2120.

[4] Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2022. GERE: Generative evidence retrieval for fact verification. In *Proc. of SIGIR.* 2184–2189.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of NAACL-HLT.* 4171–4186.

[6] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proc. ACL.* 4443–4458.

[7] Lucas Graves. 2018. Understanding the Promise and Limits of Automated Fact-checking. *Reuters Institute for the Study of Journalism* (2018).

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

[9] Ashim Gupta and Vivek Srikumar. 2021. X-Fact: A New Benchmark Dataset for Multilingual Fact Checking. In *Proc. of ACL.* 675–682.

[10] Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and S Yu Philip. 2022. CHEF: A Pilot Chinese Dataset for Evidence-Based Fact-Checking. In *Proc. of NAACL-HLT.* 3362–3376.

[11] Xuming Hu, Lijie Wen, Yusong Xu, Chenwei Zhang, and Philip Yu. 2020. SelfORE: Self-supervised Relational Feature Learning for Open Relation Extraction. In *Proc. of EMNLP.* Online, 3673–3682.

[12] Xuming Hu, Chenwei Zhang, Fukun Ma, Chenyao Liu, Lijie Wen, and S Yu Philip. 2021. Semi-supervised Relation Extraction via Incremental Meta Self-Training. In *Findings of EMNLP.* 487–496.

[13] Xuming Hu, Chenwei Zhang, Yawen Yang, Xiaohe Li, Li Lin, Lijie Wen, and S Yu Philip. 2021. Gradient Imitation Reinforcement Learning for Low Resource Relation Extraction. In *Proc. of EMNLP.* 2737–2746.

[14] Alon Jacovi and Yoav Goldberg. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, Online, 4198–4205. https://doi.org/10.18653/v1/2020.acl-main.386

[15] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proc. of ICLR.*

[16] Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing Neural Predictions. In *Proc. of EMNLP.* 107–117.

[17] Aiwei Liu, Xuming Hu, Li Lin, and Lijie Wen. 2022. Semantic Enhanced Text-to-SQL Parsing via Iteratively Learning Schema Linking Graph. In *Proc. of KDD.* 1021–1030.

[18] Aiwei Liu, Xuming Hu, Lijie Wen, and Philip S Yu. 2023. A comprehensive evaluation of ChatGPT's zero-shot Text-to-SQL capability. *arXiv preprint arXiv:2303.13547* (2023).

[19] Shuliang Liu, Xuming Hu, Chenwei Zhang, Shu'ang Li, Lijie Wen, and Philip S. Yu. 2022. HiURE: Hierarchical Exemplar Contrastive Learning for Unsupervised Relation Extraction. In *Proc. of NAACL-HLT.* 5970–5980.

[20] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[21] Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained Fact Verification with Kernel Graph Attention Network. In *Proc. of ACL.* Association for Computational Linguistics, Online, 7342–7351.

[22] Dan S Nielsen and Ryan McConville. 2022. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *Proc. of SIGIR.* 3141–3153.

[23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL.* 311–318.

[24] Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. 2021. Vera: Prediction techniques for reducing harmful misinformation in consumer health search. In *Proc. of SIGIR.* 2066–2070.

[25] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In *Proc. of EMNLP.* Association for Computational Linguistics, Copenhagen, Denmark, 2931–2937.

[26] Tal Schuster, Roei Schuster, Darsh J. Shah, and Regina Barzilay. 2020. The Limitations of Stylometry for Detecting Machine-Generated Fake News. *Computational Linguistics* 46, 2 (2020), 499–510.

[27] Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a Known Lie: Detecting Previously Fact-Checked Claims. In *Proc. of ACL.* Association for Computational Linguistics, Online, 3607–3618.

[28] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proc. of NAACL-HLT.* Association for Computational Linguistics, New Orleans, Louisiana, 809–819.

[29] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proc. NAACL-HLT.* 809–819.

[30] James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The FEVER2.0 Shared Task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER).*

[31] Nicolas Turenne. 2018. The rumour spectrum. *PloS one* 13, 1 (2018), e0189080.

[32] Nguyen Vo and Kyumin Lee. 2018. The rise of guardians: Fact-checking url recommendation to combat fake news. In *Proc. of SIGIR.* 275–284.

[33] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[34] Zhi-Xiu Ye and Zhen-Hua Ling. 2019. Multi-Level Matching and Aggregation Network for Few-Shot Relation Classification. In *Proc. of ACL.* 2872–2881.

[35] Wenpeng Yin and Dan Roth. 2018. TwoWingOS: A Two-Wing Optimization Strategy for Evidential Claim Verification. In *Proc. of EMNLP*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 105–114. https://doi.org/10.18653/v1/d18-1010

[36] Xia Zeng, Amani S Abumansour, and Arkaitz Zubiaga. 2021. Automated fact-checking: A survey. *Language and Linguistics Compass* 15, 10 (2021), e12438.

[37] Xin Zhang, Yong Jiang, Xiaobin Wang, Xuming Hu, Yueheng Sun, Pengjun Xie, and Meishan Zhang. 2022. Domain-Specific NER via Retrieving Correlated Samples. In *Proc. of COLING.* 2398–2404.