Verdict Inference with Claim and Retrieved Elements Using RoBERTa

In-Zu Gi[†] **Ting-Yu Fang**[†] **Richard Tzong-Han Tsai**[†‡*]
[†]Department of Computer Science and Information Engineering,
National Central University, Taiwan
[‡]Center for GIS, Research Center for Humanities and Social Sciences,
Academia Sinica, Taiwan
{inzugi0910, maple3142, thtsai}@g.ncu.edu.tw

Abstract

Automatic fact verification has attracted recent research attention as the increasing dissemination of disinformation on social media platforms. The FEVEROUS shared task introduces a benchmark for fact verification, in which a system is challenged to verify the given claim using the extracted evidential elements from Wikipedia documents. In this paper, we propose our 3^{rd} place three-stage system consisting of document retrieval, element retrieval, and verdict inference for the FEVEROUS shared task. By considering the context relevance in the fact extraction and verification task, our system achieves 0.29 FEVEROUS score on the development set and 0.25 FEVEROUS score on the blind test set, both outperforming the FEVEROUS baseline.

1 Introduction

The large-scale dissemination of disinformation on social media platforms intended to mislead or deceive the general population has become a major societal problem (Tan et al., 2020). For example, the widespread disinformation of the Covid-19 vaccine has caused a growth of anti-vaccination sentiment online and led to declining vaccination coverage. As the best way to stop disinformation from going viral online is early verification, recent researchers have put efforts into automatic fact verification systems.

To answer the increasing demand for such systems, the FEVER (Fact Extraction and VERification) dataset (Thorne et al., 2018) was introduced and used for the shared task of the FEVER Workshop 2018. It consists of 185,445 annotated claims with a label of "SUPPORTED", "REFUTED", or "NOT ENOUGH INFO" as well as sets of evidential sentences from the given pre-processed Wikipedia pages. Among the participated teams of the shared task, Nie et al. (2019) proposed a system consisting of three connected homogeneous networks of document retrieval, sentence selection, and claim verification. Yoneda et al. (2018) proposed a four-stage system that utilizes logistic regression models for the document retrieval and sentence retrieval stages, Enhanced Sequential Inference Model (ESIM) (Chen et al., 2017) for the natural language inference stage, and Multi-Layer Perceptron (MLP) for the aggregation stage.

To explore the ability of automatic fact verification systems over both unstructured sentences and structured table-based information, Aly et al. (2021) introduces the Fact Extraction and VERification Over Unstructured and Structured information (FEVEROUS) dataset. The shared task in 2021 uses the FEVEROUS dataset and further requires a system to be able to retrieve structured information from Wikipedia as evidence for each claim, which differs from the shared task in 2018. However, these two shared tasks still share the similar setting as a fact extraction and verification problem, which makes the pipelines and methods of the early proposed systems worth referring to. All in all, the FEVEROUS shared task in 2021 challenges a system to extract evidential elements, primarily sentences and table cells, from the given 5.4M Wikipedia documents and verify as "SUPPORTS", "REFUTES", or "NOT ENOUGH INFO" for each given claim. Systems are evaluated by jointly considering how complete the relevant Wikipedia elements are retrieved and how correct the final verification verdicts are.

In this paper, we propose a three-stage system as Figure 1 shows to improve the FEVEROUS baseline in two aspects. First, while the baseline retriever pays attention to literal relevance and word frequency with a combination method of entity matching and TF-IDF, we fine-tune the BERT model (Devlin et al., 2019) to integrate the con-

* Corresponding author.

Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER) at EMNLP 2021, pages 60–65

text relevance for finding evidential elements and downstream verdict inference. Second, the baseline predictor uses the claim and the concatenation of retrieved elements as input, having a constraint of the maximum input length. We experiment with several ways to include more elements for verdict inference. Finally, these improvements allow us to achieve substantially higher performance than the baseline.

2 System Description

Our system is a three-stage model consisting of document retrieval, element retrieval, and verdict inference. Document retrieval aims to extract the selection of the most related Wikipedia documents when only given the claim. The claim and the set of candidate elements from the most related Wikipedia documents are then given to the subsequent element retrieval to find out the most evidential elements regarding the claim. The final stage utilizes the NLI model for verdict inference, predicting the final verdict based on the most evidential elements and the claim.

2.1 Document Retrieval

Document retrieval is to extract the most related documents from 5.4M Wikipedia documents when only given the claim. A Wikipedia document is determined as related by checking if any element from the Wikipedia document is included as evidential elements for the given claim.

Our document retrieval utilizes Anserini (Yang et al., 2018), an information retrieval toolkit built on Lucene and providing an easy-to-use interface for querying. Experiments have shown that Anserini is efficient in indexing large document collections and provides modern ranking methods that are on par with research requirements.

Based on the observation that the claim often includes the title and the introductory section of the related Wikipedia document, we take the title and the first 10 elements of each Wikipedia document, normalize them by removing the links, and then build the indices of our Wikipedia document collection. We then use our Anserini to query each claim with the indices we built and retrieve k Wikipedia documents most related to the claim as well as their relatedness scores.

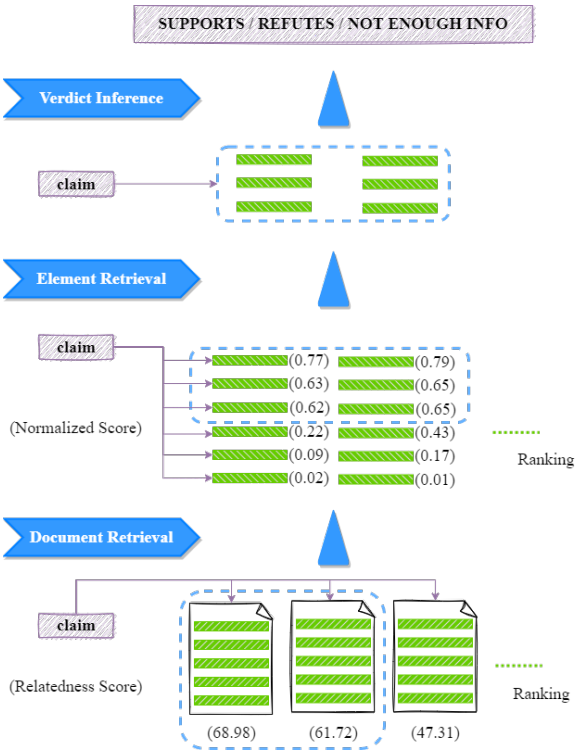

Figure 1: System Overview: Document Retrieval, Element Retrieval, and Verdict Inference.

2.2 Element Retrieval

For element retrieval, we experiment with two different approaches, the Anserini and the BERT model, to select relevant elements from documents retrieved in the previous stage. Both methods require every element to be a sequence input, including the table elements. We apply two techniques to linearize the table elements. One is converting each cell element to a sequence format of "[Header] is [Cell]." (Oguz et al., 2021), and the other is prepending the Wikipedia document title of the element in front of the converted sequence. Take the evidential cell element "Travis Hafner" as an example, we first convert it as a sequence "Player is Travis Hafner.", and then prepend the title to the sequence as "2005 Cleveland Indians season Player is Travis Hafner.".

For our Anserini in the element retrieval stage, we use every element in the entire Wikipedia documents to build the indices of another Wikipedia elements collection. Due to the mechanism of the Anserini, we first retrieve l top related Wikipedia elements when only given the claim, and apply a filter of k most related documents to utilize the benefits of the document retrieval stage and obtain the finally retrieved m elements. The afterward filter leads to the different numbers of finally re-

trieved m elements for each claim. To improve the performance, we separate the retrieval procedure for sentences and tables by building another two separated collections.

For the second approach experimented in the element retrieval stage, we fine-tune the BERT model as a binary classification with the ground truth elements as positive and the other elements as negative from the k most related Wikipedia documents retrieved by our Anserini in the document retrieval stage. Our BERT model takes the concatenation of the claim and the element sequence as input, and we use the output prediction to calculate the normalized relatedness scores. We rank the elements according to the normalized relatedness scores and select m Wikipedia elements top related to the claim. The normalized relatedness score between the claim c_i and the j-th element is computed as:

$$p(x = 1|c_i, j) = \frac{e^{p^+}}{e^{p^+} + e^{p^-}}$$

where $x \in \{0, 1\}$ indicates whether the j-th element is positive or negative, p^+ is the prediction scores for positive, p^- is the prediction scores for negative, and $p(x = 1|c_i, j)$ is the normalized score for p^+.

Our BERT model uses the Adam optimizer (Kingma and Ba, 2015) with a learning rate of $1e^{-5}$, a batch size of 16, and 1 training epochs due to the time constraint.

2.3 Verdict Inference

On the third stage of the FEVEROUS shared task, NLI is a task that matches the scenario of classifying the semantic relationship between the claim and the retrieved elements as "SUPPORTS", "REFUTES", or "NOT ENOUGH INFO" (NEI). Therefore, we adopt the RoBERTa (Liu et al., 2019) NLI model pre-trained on well-known NLI datasets, including SNLI, MNLI, FEVER-NLI, ANLI (Nie et al., 2020), and experiment on its variants with aggregation method on top of it to fully utilize the semantic information of retrieved elements in the previous stage.

For simplicity, we name our RoBERTa NLI model without aggregation as **RoBERTa**, RoBERTa NLI model with logical aggregation as **RoBERTa-LOG**, and RoBERTa NLI model with MLP for aggregation as **RoBERTa-MLP**. Both aggregation methods have been proved effective in Yoneda et al. (2018). Our RoBERTa takes the claim and the concatenation of all retrieved elements as input, while our RoBERTa-LOG and RoBERTa-MLP take the claim and each retrieved element as input.

Method	k	Recall
Baseline	5	0.69
Our Anserini	5	0.69
	10	0.73

Table 1: Recall is calculated by the frequency of the ground truth document occurrence in the retrieved documents. **k** indicates the number of retrieved documents.

The RoBERTa NLI models are fine-tuned with ground truth labels in FEVEROUS training set and additionally sampled NEI instances to get rid of the unbalanced labeling problem. We use the adam optimizer with a learning rate of $1e^{-6}$, a batch size of 8, a scheduler to watch on the development loss, and a total of 7 training epochs. Our RoBERTa-LOG simply merges the NLI predictions and outputs the label obtaining the highest point. For our RoBERTa-MLP, our MLP containing two fully connected layers and ReLU is trained jointly with the RoBERTa NLI models.

3 Results

3.1 Document Retrieval Results

To evaluate the performance of the document retrieval, we measure the recall metric and the results are shown in Table 1. Our document retriever achieves a document coverage of 69% when retrieving the top 5 documents and 73% when retrieving the top 10 documents. When deciding the value of k, it is a trade-off between retrieval performance and computational resources. As a result, we set $k = 5$ for the downstream element retrieval using the BERT model and at the same time experiment with different settings for the downstream element retrieval using the Anserini.

3.2 Element Retrieval Results

Table 2 shows the development set results using our Anserini. The retriever of $l = 5000$ with prepending achieves better performance than the retrievers without prepending. From the results, we observe that prepending title improves recall performance. Nevertheless, to meet the submission requirement of at most 5 sentences and 25 cells for each claim, the averagely obtained 58 elements require further

l	Prepend	k	avg-m	Recall
5000	-	5	24.4	0.35
50000	-	5	41.5	0.40
5000	✓	10	58	0.55
1000 (s)	✓	10	4.9	0.50
5000 (c)	✓	20	19.9	0.43

Table 2: Recall is calculated on an element level with our Anserini. **l** indicates the number of retrieved elements by our Anserini. (s) and (c) indicate separated retrieval for sentence and cells respectively. **Prepend** indicates whether the title is prepended to the linearized element sequence. **k** indicates the number of retrieved documents used to filter the elements. **avg-m** indicates the averagely retrieved m elements eventually after the filter of k most related documents on l elements.

Prepend	m	Sentence	Table	Overall
-	40	0.73	0.16	0.37
✓	40	0.71	0.52	0.59

Table 3: Coverage for sentence, table, and overall performance with our BERT using the 5 most related Wikipedia documents retrieved by our Anserini in the document retrieval stage. **m** indicates the number of retrieved elements. **Prepend** indicates whether the title is prepended to the linearized element sequence.

control on retrieved numbers of each element type. Experiments show that different combinations of l and k as well as separating the retrieval for sentences and tables respectively achieve comparable performance to the retriever of $l = 5000$ with prepending and provide better supervision of the number and type of the retrieved elements.

Table 3 shows the development set results using our BERT. We observe similar results with our Anserirni, which the retriever with prepending achieves better performance than the retrievers without prepending.

We use our Anserini of $l = 5000$ with prepending and our BERT with prepending to have a relatively fair comparison between our two approaches for element retrieval. Our Anserini of $l = 5000$ with prepending covers 55% of all elements, while our BERT with prepending covers 59%, showing our BERT substantially outperforms our Anserini.

3.3 Verdict Inference Results

Table 4 shows the development set results of our models trained on a training subset. We observe that the RoBERTa-LOG reaches 0.52 F_1 score and the RoBERTa-MLP reaches 0.22 F_1 score, both

Model	S	R	NEI	Overall
Baseline	0.88	0.85	0.30	0.67
RoBERTa	0.88	0.84	0.35	0.69
RoBERTa-LOG	0.74	0.68	0.14	0.52
RoBERTa-MLP	0.66	0.00	0.00	0.22

Table 4: Performance of different verdict inference methods trained on a training subset. Scores are reported on the development set in per-class F_1, with **S** represents "SUPPORTS", **R** represents "REFUTES", and **NEI** represents "NOT ENOUGH INFO". The overall score is reported using macro-averaged F_1.

are much lower than the RoBERTa. This indicates that, while each claim in the development set only requires an average of 4.6 elements to reach the golden truth label according to our analysis, it is inappropriate for our RoBERTa-LOG and RoBERTa-MLP to take all thirty elements evenly for each claim. Therefore, we choose to use our RoBERTa and simplify the input by removing potentially repeated words to allow more elements included for verdict inference.

We also test the performance of our RoBERTa with different element retrieval methods using the FEVEROUS scorer as shown in Table 5. The performance of evidence is reported with a restriction of at most 5 sentences and 25 cells as the FEVEROUS scorer limits. We observe that the quality of the upstream data is crucial to the performance of the downstream task, as our RoBERTa taking the elements from our Anserini and BERT reach 0.58 and 0.60 accuracy, respectively. Experiments also show that our RoBERTa taking the elements from our BERT is more robust than our Anserini with improvements in evidence precision, recall, and F_1 score. Our RoBERTa taking the elements from our BERT also outperforms the baseline with 0.1 improvements on FEVEROUS score.

According to our observation of the performance on all three stages, we decide our final system as the combination of our Anserini for document retrieval, our BERT for element retrieval, and our RoBERTa for verdict inference. The results of the blind test of our final system are presented in Table 6. Our final system is proved robust and outperforms the FEVEROUS baseline.

4 Error Analysis

Our system proves that performing document retrieval, element retrieval and verdict inference in the three-phase procedure is a proper pipeline for

System	FEVEROUS	Accuracy	Evidence		
			Precision	Recall	F_1
Baseline	0.19	0.53	0.12	0.30	0.17
Ground Truth+RoBERTa	0.84	0.84	1.00	0.99	0.99
Anserini+Anserini+RoBERTa	0.22	0.58	0.09	0.32	0.14
Anserini+BERT+RoBERTa	0.29	0.60	0.10	0.42	0.17

Table 5: Performance of different element retrieval methods using our RoBERTa. Scores are reported on the development set using the FEVEROUS scorer. The Anserini uses $l = 5000$ with prepending and a filter of $k = 10$ documents. The BERT uses $k = 5$ documents retrieved from the previous stage and utilizes prepending for element retrieval.

System	FEVEROUS	Accuracy	Evidence		
			Precision	Recall	F_1
Baseline	0.1773	0.4548	0.1017	0.2878	0.1503
Anserini+Anserini+RoBERTa	0.2215	0.5108	0.0810	0.3414	0.1310
Anserini+BERT+RoBERTa (final)	**0.2514**	**0.5229**	0.0991	0.3907	0.1581

Table 6: Performance of systems on blind test results. Our final system is Anserini for document retrieval with $k = 5$, BERT for element retrieval with prepending, and RoBERTa for verdict inference. The candidate system is Anserini for both document and separated retrieval of different element types as well as RoBERTa for verdict inference.

the fact extraction and verification shared task. For evidence retrieval (document and element retrieval), our proposed methods of the BM25-based Anserini and the context-aware BERT model considers both the presence of certain keywords and semantic context. Hence, it is able to extract related elements over unstructured sentences and structured table cells.

Nevertheless, the retrieval performance still has room for improvement in two aspects. One is to tune the balance weighting between the presence of certain keywords and the semantic context. The other is to attentively design the fine-tuning of the BERT model for element retrieval. During the fine-tuning process, we use the output from the document retrieval stage and set as negative labels for all elements from the 5 most related documents that are not annotated as evidence of the corresponding claim. While each claim has only a maximum of 3 evidence sets and an average of nearly 4 elements per set, our BERT model suffers from the unbalanced labeling problem, in which the process includes massive negative and few positive instances. The positive instances in the fine-tuning are also rather few because we use the output of the document retrieval stage that only retrieves a document coverage of 69%.

5 Conclusion

We describe our 3^{rd} place system for the FEVEROUS shared task via the three-stage setup of document retrieval, element retrieval, and verdict inference. By considering the context relevance in the fact extraction and verification task, our system achieves 0.29 FEVEROUS score on the development set and 0.25 FEVEROUS score on the blind test set, both outperforming the FEVEROUS baseline.

References

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. *CoRR*, abs/2106.05707.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN,*

USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 4171–4186. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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. *CoRR*, abs/1907.11692.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6859–6866. AAAI Press.

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*. Association for Computational Linguistics.

Barlas Oguz, Xilun Chen, Vladimir Karpukhin, Stan Peshterliev, Dmytro Okhonko, Michael Schlichtkrull, Sonal Gupta, Yashar Mehdad, and Scott Yih. 2021. Unik-qa: Unified representations of structured and unstructured knowledge for open-domain question answering.

Reuben Tan, Bryan Plummer, and Kate Saenko. 2020. Detecting cross-modal inconsistency to defend against neural fake news. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2081–2106, 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, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 809–819. Association for Computational Linguistics.

Peilin Yang, Hui Fang, and Jimmy Lin. 2018. Anserini: Reproducible ranking baselines using lucene. *ACM J. Data Inf. Qual.*, 10(4):16:1–16:20.

Takuma Yoneda, Jeff Mitchell, Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. UCL machine reading group: Four factor framework for fact finding (HexaF). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 97–102, Brussels, Belgium. Association for Computational Linguistics.