Adapting Open Domain Fact Extraction and Verification to COVID-FACT through In-Domain Language Modeling

Zhenghao Liu[1,2], **Chenyan Xiong**[5], **Zhuyun Dai**[6], **Si Sun**[4], **Maosong Sun**[1,3], **Zhiyuan Liu**[1,3]
[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China
Institute for Artificial Intelligence, Tsinghua University, Beijing, China
Beijing National Research Center for Information Science and Technology
[2]State Key Lab on Intelligent Technology and Systems, Tsinghua University, Beijing, China
[3]Beijing Academy of Artificial Intelligence
[4]Department of Electronic Engineering, Tsinghua University, Beijing, China
[5]Microsoft Research, Redmond, USA
[6]Carnegie Mellon University, USA

Abstract

With the epidemic of COVID-19, verifying the scientifically false online information, such as fake news and maliciously fabricated statements, has become crucial. However, the lack of training data in the scientific domain limits the performance of fact verification models. This paper proposes an in-domain language modeling method for fact extraction and verification systems. We come up with SciKGAT to combine the advantages of open-domain literature search, state-of-the-art fact verification systems and in-domain medical knowledge through language modeling. Our experiments on SCIFACT, a dataset of expert-written scientific fact verification, show that SciKGAT achieves 30% absolute improvement on precision. Our analyses show that such improvement thrives from our in-domain language model by picking up more related evidence pieces and accurate fact verification. Our codes and data are released via Github[1].

1 Introduction

Online contents with false information, such as lies, rumors and conspiracy theories, have been growing significantly and spreading widely during the COVID-19 epidemic. An automatic fact-checking system is urgently needed to check these scientific claims, which can avoid undesired consequences. Automatic fact-checking has drawn lots of attention from NLP community. Researchers mainly focus on stopping misinformation transmission through videos and texts (Cinelli et al., 2020; Hossain et al., 2020; Li et al., 2020; Serrano et al., 2020).

The scientific fact verification task (Wadden et al., 2020) is come up to deal with COVID-FACT with high-quality articles of spanning domains from basic science to clinical medicine. Nevertheless, the small-scale training data of SCIFACT may limit the performance of COVID-FACT checking. The state-of-the-art model (Wadden et al., 2020) achieves only 46.6% precision of fact verification, which is hard to be trusted for users.

This paper presents the Scientific KGAT (SciKGAT) to deal with low-resource COVID-FACT verification. SciKGAT employs the in-domain language model in the fact extraction and verification pipeline (Thorne et al., 2018; Wadden et al., 2020) to adapt fact-checking into COVID domain. The in-domain language model transfers COVID domain knowledge into pre-trained language models with continuous training and learns medical token semantics towards COVID with mask language model based training. The state-of-the-art fact verification model KGAT (Liu et al., 2020; Ye et al., 2020) is also used in SciKGAT for multi-evidence reasoning in the fact verification module.

Our experiments show that the in-domain language modelings achieve better performance for various components in the whole fact extraction and verification pipeline by achieving more accurate evidence selection and fact verification. Our in-domain language modelings improve the fact verification performance with more than 10% absolute F_1 score and 30% absolute precision (from 46.6% to 76%) than previous state-of-the-art on SCIFACT. Such improvement shows that our model provides a set of solutions for low-resource fact verification tasks, such as COVID-19.

2 Related Work

Existing fact extraction and verification models usually employ a three-step pipeline system (Chen et al., 2017): document retrieval (abstract retrieval), sentence selection (rationale selection) and fact verification (Thorne et al., 2018; Wadden et al., 2020).

The preliminary fact verification methods concatenate all evidence pieces (Nie et al., 2019; Wad-

[1] `https://github.com/thunlp/KernelGAT`

November 16 - 20, 2020. ©2020 Association for Computational Linguistics

den et al., 2020) for fact verification. KGAT (Liu et al., 2020) conducts fine-grained multiple evidence reasoning with a graph and achieves the state-of-the-art for fact verification (Ye et al., 2020).

The reasoning ability of the pre-trained language model is crucial and helps improve fact verification performance (Devlin et al., 2019; Li et al., 2019; Zhou et al., 2019; Soleimani et al., 2019). Some work (Beltagy et al., 2019; Lee et al., 2020) transfers medical domain knowledge into pre-trained language models for better medical semantic understanding, which provides a potential way to deal with COVID-FACT checking problem.

3 Methodology

This section describes our SciKGAT for fact extraction and verification. We first introduce the pipeline of fact extraction and verification (Sec. 3.1) and then continuously train the BERT based model (Sec. 3.2) for the whole.

3.1 Preliminary

Given a claim c, we aim to predict the claim label y. We usually implement the fact extraction and verification pipeline with three steps: abstract retrieval, rationale selection and fact verification.

Abstract Retrieval. For the claim c and abstract $D = \{a_1, \dots, a_l\}$, we aim to retrieve three abstracts for the following steps.

We first retrieve top-100 abstracts with TF-IDF from the abstract collection D, which is the same as the previous work (Wadden et al., 2020). For the claim c and abstract abstract $a = \{e_1, \dots, e_k\}$ with k evidence pieces and title t, we concatenate claim, title and abstract to get the representation $\mathcal{H}^e$ of the pair $\langle c, a\rangle$ with BERT (Devlin et al., 2019):

$$\mathcal{H} = \text{BERT}(\text{[CLS]} \circ c \circ \text{[SEP]} \circ t \circ a \circ \text{[SEP]}), \quad (1)$$

where $\circ$ is the concatenate operation. The representation $\mathcal{H}$ of $\langle c, a\rangle$ consists of representations of tokens from both claim and evidence. The 0-th representation $\mathcal{H}_0$ denotes the `[CLS]` representation. The relevance label y_a between claim c and abstract a is calculated:

$$p(y_a|c, a) = \text{softmax}_{y_a}(\text{MLP}(\mathcal{H}_0)). \quad (2)$$

We rerank abstracts according to the probability $p(y_a = 1|c, a)$ and top-3 abstracts are reserved.

Rationale Selection. Given the retrieved abstract a, rationale selection focuses on selecting relevant sentences for fact verification.

Similarly, for the evidence e of the retrieved abstract a, we can get the representation H of claim and evidence pair $\langle c, e\rangle$:

$$H = \text{BERT}(\text{[CLS]} \circ c \circ \text{[SEP]} \circ e \circ \text{[SEP]}). \quad (3)$$

Then we predict the relevance label y_r of claim c and evidence e:

$$p(y_r|c, e) = \text{softmax}_{y_r}(\text{MLP}(H_0)). \quad (4)$$

The related evidence pieces ($p(y_r = 0|c, e) < p(y_r = 1|c, e)$) are reserved to form the retrieved evidence set $E = \{e_1, ..., e_q\}$ of each abstract a.

Fact Verification. For the claim c and retrieved evidence set E, fact verification model aims to predict claim label y. We employ the state-of-the-art model KGAT (Liu et al., 2020) as our fact verification module. For the i-th evidence e_i in the evidence set E, we can get the sentence pair representation H^i of the i-th pair $\langle c, e_i\rangle$ through BERT. Then the probability of claim label y is calculated:

$$p(y|c, E) = \text{KGAT}(H^1, \dots, H^q). \quad (5)$$

3.2 Continuous In-Domain Training

To deal with the low-resource COVID-FACT checking, we propose continuous training methods to transfer domain knowledge into pretrained language models.

For COVID-FACT checking, the medical domain knowledge is useful to understand medical words (Beltagy et al., 2019). However, these medical domain pre-trained language models will be out-of-date with the medical development or emergence of a new virus, such as COVID-19.

Continuous in-domain training provides a potential way to deal with this problem with the latest medical corpus. Hence we come up with two in-domain language models for the fact extraction and verification pipeline with continuous training.

Rationale prediction based training. We first come up with the rationale prediction style training to continuously train BERT for better reasoning ability towards the COVID-FACT. For the claim and evidence $\langle c, e\rangle$, we optimize BERT model with supervisions from SCIFACT:

$$L_r(c, e) = \text{CrossEntropy}(p(y_r|c, e), y_r^*), \quad (6)$$

where y_r^* denotes the ground truth rationale prediction label of the pair $\langle c, e\rangle$. Then we get a supervised in-domain language model, BERT-RP, for the fact verification module.

Model	Development set						Testing Set					
	Sentence Level			Abstract Level			Sentence Level			Abstract Level		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Baselines												
SciBERT	45.78	38.52	41.84	51.93	44.98	48.21	-	-	-	-	-	-
RoBERTa	46.51	38.25	41.98	53.30	46.41	49.62	38.6	40.5	39.5	46.6	46.4	46.5
SciKGAT												
KGAT	57.07	31.97	40.98	72.73	38.28	50.16	-	-	-	-	-	-
SciKGAT (w. A)	42.07	**47.81**	44.76	47.66	**58.37**	52.47	40.50	**48.38**	44.09	47.06	**57.66**	51.82
SciKGAT (w. AR)	50.00	**47.81**	48.88	53.15	56.46	54.76	41.67	45.95	43.70	47.47	54.96	50.94
SciKGAT (Full)	**74.36**	39.62	**51.69**	**84.26**	43.54	**57.41**	**61.15**	42.97	**50.48**	**76.09**	47.30	**58.33**

Table 1: Overall Performance of Fact Extraction and Verification. RoBERTa is the large version. SciKGAT (w. A) and SciKGAT (w. AR) are ablation models with the abstract retrieval and evidence selection of SciKGAT.

Mask language model based training. To help the model better comprehend the semantics of COVID related words, we substitute tokens with `[MASK]` and ask the model to generate appropriate tokens for filling it. With continuous training, the language model now sees the language from the new corpus, thus being able to pick up the new terminologies, such as COVID-19. The continuous training with COVID related corpus is able to better capture the context/semantics of such new terminologies (Gururangan et al., 2020).

We use data from COVID-19 Open Research Dataset Challenge[2] for continuous training, which towards the medical topic. In this corpus, there are about 86K papers before 2020, which are about coronaviruses but not about COVID-19, and 54K papers after 2020. Based on the filters used by AI2 to create this dataset, those papers that after 2020 are almost about COVID-19. Thus roughly there are about 40% papers in this corpus that are about COVID-19 (Wang et al., 2020).

4 Experimental Methodology

This section describes the dataset, evaluation metrics, baselines, and implementation details.

Dataset. The recently released dataset SCIFACT (Wadden et al., 2020) is leveraged in our experiments. It consists of 1,409 annotated claims with 5,183 scientific articles. All claims are classified as SUPPORT, CONTRADICT or NOT ENOUGH INFO. The training, development and testing sets contain 809, 300 and 300 claims, respectively. FEVER (Thorne et al., 2018) is also used by official baselines to train the fact verification modules of baselines and our models. The FEVER consists of 185,455 annotated claims with 5,416,537 Wikipedia documents.

Evaluation Metrics. Precision, Recall and F_1 score are used to evaluate model performance, following SCIFACT (Wadden et al., 2020). These evaluations are inspired by FEVER score (Thorne et al., 2018) and consider if the evidence is selected correctly from the abstract level and sentence level.

Baselines. Since the scientific fact verification task is recently released, our baselines are mainly from Wadden et al. (2020). They first use TF-IDF for abstract retrieval and then use RoBERTa (Large) and SiBERT for rationale selection. KGAT and RoBERTa (Large) are leveraged for fact verification. The rationale selection module is trained with SCIFACT and the fact verification module is trained with data from FEVER and SCIFACT (Wadden et al., 2020).

Implementation Details. In all experiments, we use SciBERT, RoBERTa (Base) and RoBERTa (Large) (Liu et al., 2019; Beltagy et al., 2019), and inherit huggingface's PyTorch implementation[3]. Adam is utilized for parameter optimization. For rationale selection, we keep the same setting as Wadden et al. (2020). For abstract retrieval and fact verification, we set the max length to 256, learning rate to 2e-5, batch size to 8 and accumulate step to 4 during training. The other parameters are kept the same with KGAT (Liu et al., 2020).

For the abstract retrieval module, we follow the previous work (MacAvaney et al., 2020) and finetune our in-domain language model with the medical corpus from MS-MARCO (Bajaj et al., 2016) to fit our abstract retrieval module to the open-domain COVID related literature search.

5 Evaluation Result

This section first tests the overall performance of SciKGAT. Then it studies the impacts of our in-domain language modeling techniques in knowl-

[2] https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

[3] https://github.com/huggingface/pytorch-transformers

Ablation	Model	Evidence Retrieval			Fact Checking					
		Ranking Accuracy			Sentence Level			Abstract Level		
		Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Abstract Retrieval	TF-IDF	16.11	69.38	26.15	46.51	38.25	41.98	53.30	46.41	49.62
	w. SciBERT	19.78	85.17	32.10	**42.09**	47.27	44.53	**48.18**	56.94	52.19
	w. SciBERT-MLM	**20.33**	**87.56**	**33.00**	42.07	**47.81**	**44.76**	47.66	**58.37**	**52.47**
Rationale Selection	SciBERT	36.90	**65.03**	47.08	43.22	46.99	45.03	48.94	55.02	51.80
	SciBERT-MLM	**43.73**	60.93	**50.91**	**50.00**	**47.81**	**48.88**	**53.15**	**56.46**	**54.76**
Fact Verification	SciBERT	43.73	60.93	50.91	36.55	**38.25**	37.38	36.92	**45.93**	40.94
	w. KGAT	-	-	-	51.61	34.97	41.69	58.99	39.23	47.13
	w. KGAT (RP Init)	-	-	-	**60.10**	33.33	**42.88**	**66.38**	36.84	47.38
	w. KGAT (MLM Init)	-	-	-	56.00	34.43	42.64	65.32	38.76	**48.65**
	RoBERTa-Base	43.73	60.93	50.91	42.72	36.89	39.59	44.50	**46.41**	45.43
	w. KGAT	-	-	-	61.05	31.69	41.73	**68.87**	34.93	46.35
	w. KGAT (RP Init)	-	-	-	**61.19**	36.61	45.81	67.48	39.71	50.00
	w. KGAT (MLM Init)	-	-	-	60.35	**37.43**	**46.21**	67.19	41.15	**51.04**
	RoBERTa-Large	43.73	60.93	50.91	50.00	**47.81**	48.88	53.15	**56.46**	54.76
	w. KGAT	-	-	-	62.87	40.71	49.42	72.39	46.41	56.56
	w. KGAT (RP Init)	-	-	-	73.47	39.34	51.25	83.33	43.06	56.78
	w. KGAT (MLM Init)	-	-	-	**74.36**	39.62	**51.69**	**84.26**	43.54	**57.41**

Table 2: In-Domain Language Model Performance of Fact Extraction and Verification on Development Set. Model performance with SciBERT on both abstract retrieval and rationale selection scenarios is presented. For fact verification, the in-domain language modeling methods, MLM (Mask Language Model) and RP (Rationale Prediction), are evaluated with the state-of-the-art fact verification model KGAT (Liu et al., 2020; Ye et al., 2020).

Claim: Basophils counteract disease development in patients with systemic lupus erythematosus (SLE).
Evidence 1: ... ***basophils*** and IgE autoantibodies amplify autoantibody production that ***leads to lupus nephritis*** ... **Evidence 2:** ***Individuals with SLE also have elevated*** serum IgE, self-reactive IgEs and ***activated basophils*** that ... **SciKGAT:** Contradict **RoBERTa:** Not Enough Info
Claim: In adult tissue, most T cells are memory T cells.
Evidence 1: ***Whereas adult tissues contain a predominance of memory T cells***, in pediatric blood and tissues the main subset consists of naive recent thymic emigrants ... **SciKGAT:** Support **KGAT:** Contradict

Table 3: Examples of Fact Verification. All models are implemented with RoBERTa (Large). The contents are ***emphasized*** that can verify the given claim.

edge transfer. Finally, it provides case studies.

5.1 Overall Performance

The overall performance of SciKGAT is shown in Table 1. The official baseline model uses TF-IDF for abstract retrieval and RoBERTa (Large) for rationale selection and fact verification, which is state-of-the-art. We add modules of SciKGAT step by step to evaluate the model's effectiveness.

SciKGAT (w. A) and SciKGAT (w. AR) show significant improvement than baselines, which demonstrates our literature search with an in-domain language model is effective in selecting related evidence from abstract and sentence levels. For fact verification, our SciKGAT improves pipeline performance by achieving 30% improvement on label prediction precision. The high precision of fact verification demonstrates that our model has the ability to provide high quality and convinced COVID-FACT verification results.

5.2 In-Domain Effectiveness

In this experiment, we evaluate the impacts of the in-domain language model on individual fact extraction and verification components of SciKGAT.

As shown in Table 2, we first compare SciBERT and SciBERT-MLM on the abstract retrieval and rationale selection tasks. Then we fix the selected evidence and evaluate the reasoning ability of the fact verification module, using two kinds of in-domain language models, MLM model (*mask language model training*) and RP model (*rationale prediction training*) with three BERT variants.

For abstract retrieval and rationale selection, SciBERT-MLM shows better ranking accuracy than SciBERT, and consequently results in better fact verification results. It demonstrates that the mask language model learns specific medical domain knowledge through the latest COVID related papers and thrives on our evidence selection parts with continuous training.

Then we evaluate the effectiveness of in-domain language models on fact verification with various BERT based models. Our in-domain language models significantly improve fact verification performance and illustrate their stronger reasoning ability compared to vanilla pre-trained language models. Compare to the RP model, MLM model usually achieves better performance. Importantly, MLM model does not rely on annotation data, providing a common resolution for COVID related tasks. The consistent improvement on all BERT variants further manifests the robustness of our model.

5.3 Case Study

As shown in Table 3, two examples from the development set are used to illustrate SciKGAT's effectiveness for fact verification.

In the first example, both evidence 1 and evidence 2 indicate that *basophils* can lead to *systemic lupus erythematosus*, which contradicts the claim. The concatenation based model, RoBERTa, fails to verify the claim, while SciKGAT makes the right prediction. It demonstrates the effectiveness of KGAT's fine-grained reasoning with multiple evidence pieces. In the second example, the evidence piece indicates that *memory T cells* are the most in *T cells* for adults. SciKGAT predicts claim label correctly and shows its effectiveness by recognizing and comprehending these medical phrases, which thanks to the in-domain language modeling.

6 Conclusion

This paper presents in-domain language modeling methods for open domain fact extraction and verification, which transfer domain knowledge for the COVID-FACT checking task. Our experiments show that our pipeline significantly improves the fact-checking performance of the state-of-the-art model with more than 30% absolute prediction precision. Our analyses illustrate that our model has stronger reasoning ability with continuous training and benefits from COVID related knowledge.

Acknowledgments

This work is supported by the National Key RD Program of China (2020AAA0105200), Beijing Academy of Artificial Intelligence (BAAI) and the NExT++ project from the National Research Foundation, Prime Minister's Office, Singapore under its IRC@Singapore Funding Initiative.

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. Scibert: Pretrained contextualized embeddings for scientific text. *arXiv preprint arXiv:1903.10676*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of ACL*, pages 1870–1879.

Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. The covid-19 social media infodemic. *arXiv preprint arXiv:2003.05004*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*, pages 8342–8360.

Tamanna Hossain, Robert L Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sameer Singh, and Sean Young. 2020. Detecting covid-19 misinformation on social media.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Tianda Li, Xiaodan Zhu, Quan Liu, Qian Chen, Zhigang Chen, and Si Wei. 2019. Several experiments on investigating pretraining and knowledge-enhanced models for natural language inference. *arXiv preprint arXiv:1904.12104*.

Yunyao Li, Tyrone Grandison, Patricia Silveyra, Ali Douraghy, Xinyu Guan, Thomas Kieselbach, Chengkai Li, and Haiqi Zhang. 2020. Jennifer for covid-19: An nlp-powered chatbot built for the people and by the people to combat misinformation.

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*.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In *Proceedings of ACL*, pages 7342–7351.

Sean MacAvaney, Arman Cohan, and Nazli Goharian. 2020. Sledge: A simple yet effective baseline for coronavirus scientific knowledge search. *arXiv preprint arXiv:2005.02365*.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of AAAI*, pages 6859–6866.

Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. 2020. Nlp-based feature extraction for the detection of covid-19 misinformation videos on youtube.

Amir Soleimani, Christof Monz, and Marcel Worring. 2019. BERT for evidence retrieval and claim verification. *arXiv preprint arXiv:1910.02655*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of NAACL*, pages 809–819.

David Wadden, Kyle Lo, Lucy Lu Wang, Shanchuan Lin, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. 2020. Cord-19: The covid-19 open research dataset. *ArXiv*.

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential reasoning learning for language representation. *arXiv preprint arXiv:2004.06870*.

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 ACL*, pages 892–901.