Improving Evidence Retrieval for Automated Explainable Fact-Checking

Chris Samarinas[1], Wynne Hsu[2,3], and Mong Li Lee[2,3]

[1]Institute of Data Science, National University of Singapore
[2]NUS Centre for Trusted Internet and Community
[3]School of Computing, National University of Singapore

Abstract

Automated fact-checking on a large-scale is a challenging task that has not been studied systematically until recently. Large noisy document collections like the web or news articles make the task more difficult. We examine the performance of a three-stage automated fact-checking system using various evidence retrieval and selection methods. We demonstrate that hybrid passage retrieval using sparse and dense representations leads to much higher evidence recall in a noisy setting. We also propose two sentence selection approaches, an embedding-based selection using a dense retrieval model, and a sequence labeling approach for context-aware selection. The embedding-based selection achieves very high recall across two different datasets, while the sequence labeling model achieves higher precision and improves the verification accuracy compared to context-agnostic sentence selection approaches. Using the same three-stage architecture, we built Quin, a large-scale fact-checking system for the COVID-19 pandemic.

1 Introduction

Nowadays, there are numerous organizations from many different countries dedicated to verifying facts online. However, manual fact-checking is very time consuming and intractable on a large scale in the presence of social media and individual misinformation spreaders. The automated early detection of false claims and the existence of tools that can assist in their faster verification are very critical for the minimization of the negative social impact. The recent introduction of the FEVER dataset (Thorne et al., 2018) allowed for the first time the systematic evaluation of automated fact-checking systems. However, this dataset does not reflect very well the nature of the noisy data a real-world fact-checking system has to process. The currently proposed systems have limited actual applications due to the big challenge in the retrieval of relevant evidence from thousands of sources and millions of documents. Motivated by this, in this work we examine how we can make the retrieval of relevant evidence for automated verification of claims more tractable on a large scale. We introduce the Factual-NLI+ dataset, an extension of the FEVER dataset with synthetic examples from question answering datasets and noise passages from web search results, and examine how dense representations can improve the first-stage retrieval recall of passages for fact-checking in a noisy setting. We also experiment with sentence selection approaches following the passage retrieval, and show that context-aware sentence selection with a sequence labeling model improves the verification accuracy, while an embedding-based selection method achieves very high recall across different datasets. We deployed a demo of our fact-checking architecture for the verification of claims about the COVID-19 pandemic using news articles and research papers.

2 Related Work

Automated claim verification using a large corpus is a complex task, that had not been studied systematically until recently. The recently introduced Fact Extraction and VERification dataset (FEVER) (Thorne et al., 2018) helped with the systematic evaluation of various systems. This dataset contains claims that are supported or refuted by specific evidence from Wikipedia articles, while some of them do not have any relevant evidence. The top performing system for the FEVER task, relies on sparse passage retrieval in the first stage and a neural semantic matching network for second stage passage filtering, evidence sentence selection and claim verification (Nie et al.,

2019). The second performing system (Yoneda et al., 2018), uses a naive approach for document retrieval and sentence selection. The document retrieval in the first stage relies on the mentions of entities in the claim, and in a second stage a logistic regression model, using features like keyword matches with the first sentence and the rest of the article, decides which article most likely contains the gold evidence. A similar regression model is used for sentence selection, and eventually the claim verification is done using a fine-tuned Enhanced Sequential Inference Model (ESIM) (Chen et al., 2017). The third system in the leaderboard, relies on a basic entity linking approach for document retrieval, a fine-tuned standard ESIM model (Chen et al., 2017) for sentence selection and a slightly modified ESIM model for entailment classification (Hanselowski et al., 2018).

Apart from document retrieval, sentence (evidence) selection is also a very critical task in fact-checking. There is some recent work that attempted to solve this specific task. (Atanasova et al., 2020) proposed a model for extracting relevant evidence sentences from multi-sentence passages in the LIAR-PLUS dataset (Alhindi et al., 2018). The proposed BERT-based model takes as input a set of sentences, and in the output gives the ids of the sentences that should be selected as evidence. They also observed that joint training on veracity prediction and explanation generation performs better than training individual models for these tasks. (Stammbach and Ash) investigated how the few-shot learning capabilities of the GPT-3 model (Brown et al., 2020) can be used for generating fact-checking explanations.

The success of transformer-based (Vaswani et al., 2017) language models like BERT (Devlin et al., 2019), lead to significant improvements in passage retrieval for open-domain question answering (Lee et al., 2019; Karpukhin et al., 2020; Xiong et al., 2020). Passage re-ranking with cross-attention models based on BERT (Nogueira et al., 2019) and its variants improved the results returned by sparse retrieval, and dense encoders even gave better results than sparse tfidf retrieval (Chang et al., 2020; Karpukhin et al., 2020; Samarinas et al., 2020). Retrieval based on sparse representations and exact keyword matching can be very restrictive for various queries, and dense representations mitigate this problem. While dense retrieval has been very successful in open-domain question answering, its benefit has not been widely studied yet in fact-checking.

Natural Language Inference, also known as textual entailment classification, is the task of detecting whether a hypothesis statement is entailed by a premise passage. It is essentialy a text classification problem, where the input is a pair of premise-hypothesis (P, H) and the output a label $y \in \{\text{entailment, contradiction, neutral}\}$. Datasets like the Stanford Natural Language Inference corpus (SNLI) (Bowman et al., 2015) and the Multi-Genre Natural Language Inference corpus (MultiNLI) (Williams et al., 2018) helped with the development of models for this task. An NLI model is often a core component of many automated fact-checking systems, therefore improving its accuracy is critical for the overall performance of such systems.

3 Methodology

Our automated fact-checking system, named Quin, has a three-step pipeline for verifying a given claim: passage retrieval from a corpus, sentence selection and entailment classification (figure 1). In this section, we give a formal definition of the claim verification task, we describe the datasets we used and the approaches we tested for passage retrieval, sentence selection and entailment classification.

3.1 Problem Formalization

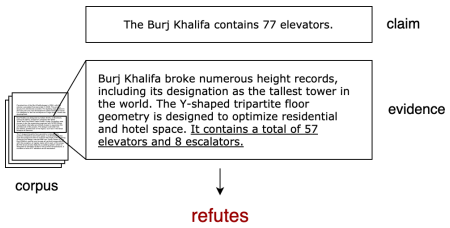

Figure 1: The claim verification task.

The automated claim verification task is defined as follows: given a textual claim c_m and a corpus $D = \{d_1, d_2, ..., d_n\}$, where every passage $d_i = \{s_1^i, s_2^i, ..., s_3^i\}$ is comprised of sentences s_j^i, a system must return a set of evidence sentences $\hat{E_m} = \{s_1, s_2, ...s_k\} \subset \bigcup d_i$ and a label $\hat{y_m} \in \{\text{supports, refutes, not enough info}\}$. Given a golden set of claims $c_m \in C$, relevant evidence E_m and labels y_m, the system has correct output if $\hat{y_m} = y_m$ and $E_m \subseteq \hat{E_m}$. This is the definition of the FEVER task, and the accuracy de-

scribed is known as FEVER score (Thorne et al., 2018). In this work (like in all the other proposed systems), we split the claim verification task into three sub-tasks; *passage retrieval*, *sentence selection* and *entailment classification* (as seen in figure 1). We evaluate every task individually and also perform evaluation of the whole pipeline.

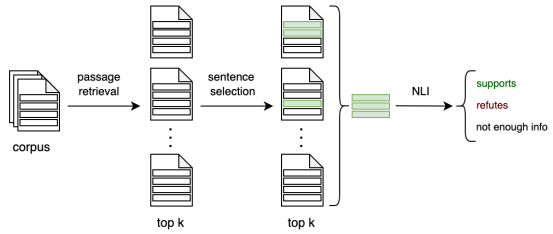

Figure 2: The three-stage verification process.

3.2 Datasets

FEVER The FEVER dataset is the first and most popular to date benchmark for evaluation of fact-checking systems. It consists of claims that can be supported or refuted by evidence from Wikipedia articles, including also claims that cannot be verified based on the Wikipedia passages. Systems are evaluated for their ability to retrieve the correct passages and evidence sentences and give correct label (supports, refutes, not enough info). FEVER has two major issues. First of all, it contains bias (Schuster et al., 2019). More specifically, the claims contain annotation artifacts that encourage the trained verification models to ignore the evidence and make predictions based only on the claim artifacts. Schuster et. al proposed a method to mitigate this bias, leading however to much lower verification accuracy on a new symmetric test set. Secondly, the passage retrieval cannot be properly evaluated, because around 90% of the claims are about specific entities that often have only a few very relevant Wikipedia passages. In a large-scale fact-checking system, like one that relies on news articles or scientific publications, there is a lot of information from multiple sources about the entities mentioned in a given claim. Consequently, the first stage document retrieval is very important and naive keyword matching approaches would not work that well. For these reasons, we used an extension of FEVER for our experiments.

Factual-NLI Factual-NLI is a recently proposed dataset (Samarinas et al., 2020) that extends FEVER with more diverse synthetic examples derived from question answering datasets; 359,190 new entailed claims with evidence and additional contradicted claims from a rule-based approach. It was proven sufficient for training of a dense retrieval model that outperformed the traditional BM25 in passage retrieval for fact-checking. In this work, we used Factual-NLI (excluding the noisy examples from the rule-based augmentation) for the evaluation of sentence selection, entailment classification and end-to-end claim verification.

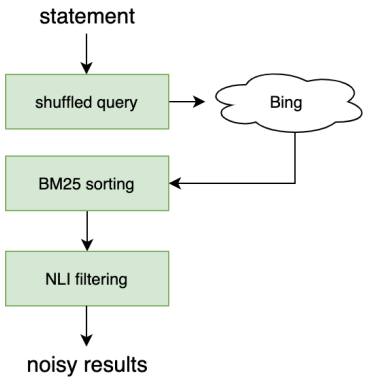

Figure 3: The augmentation process of Factual-NLI.

Factual-NLI+ For the evaluation of the passage retrieval task, we introduce a new large-scale *noisy* version of Factual-NLI. First of all, we include all the 5M Wikipedia passages used in the FEVER task. In addition, we add 'noise' passages with an augmentation process (as seen in figure 3). For every claim c_m from FEVER, we retrieved the top web results from the Bing search engine and kept the passages with the highest BM25 score that are classified as neutral by a Natural Language Inference Model (described in 3.5) with respect to c_m. For claims generated from MSMARCO queries, we included the irrelevant passages already existing for those queries. This way we included 418,650 additional passages that make the keyword-based retrieval even more challenging than FEVER and Factual-NLI. The new dataset reflects better the nature of a large-scale corpus that would be used by real-world fact-checking system that relies on multiple sources.

3.3 Passage Retrieval

For the retrieval of relevant passages, we tested the traditional *sparse* BM25 retrieval and QR-BERT, a *dense* retrieval model based on BERT (Samarinas et al., 2020). QR-BERT creates dense vectors for

passages by calculating their average token embedding. The relevance of a passage d to a claim c is then given by the dot product of their dense vectors:

$$r(c,d) = \phi(c)^T \phi(d) \quad (1)$$

Dot product search can run efficiently using an Approximate Nearest Neighbors index. In our system we created one using the FAISS library (Johnson et al., 2017). QR-BERT was trained on the Factual-NLI dataset by maximizing the sampled softmax loss:

$$L_\theta = \sum_{(c,d)\in D_B^+} r_\theta(c,d) - log\Big(\sum_{d_i \in D_B} e^{r_\theta(c,d_i)}\Big) \quad (2)$$

where D_B is the set of passages in a training batch B, and D_B^+ is the set of positive claim-passage pairs in B. It was shown that QR-BERT outperformed BM25 on a small-scale benchmark using the development set of Factual-NLI. Here, we make the comparison on a large-scale setting with the noisy extension of Factual-NLI, and examine how the dense model improves the overall performance of a fact-checking system.

3.4 Sentence Selection

We experiment with three approaches for the selection of evidence sentences from passages; a common approach with sentence-level entailment classification and two new methods.

Sentence Entailment Classification

For a given claim c, using a Natural Language Inference model (see 3.5) we select the sentences s_j from a given passage $d = \{s_1, s_2, ..., s_k\}$ so that NLI$(s_j, c) \neq$ neutral.

Embedding-based Selection

This new approach relies on the dense representations learned by the dense passage retrieval model QR-BERT. For a given claim c, using the QR-BERT model (1), we select the sentences s_j from a given passage $d = \{s_1, s_2, ..., s_k\}$ so that $r(c,d) > \lambda$ where λ a threshold defined experimentally.

Sequence Labeling

The second new approach we propose, formulates the sentence selection task as a sequence labeling problem. We adopt the BIO tagging format so that all the irrelevant tokens are classified as *O*, the first token of an evidence sentence classified as *B-EVIDENCE* and the rest tokens of an evidence sentence as *I-EVIDENCE*. We trained a model based on RoBERTa-large (Liu et al., 2019), minimizing the cross-entropy loss:

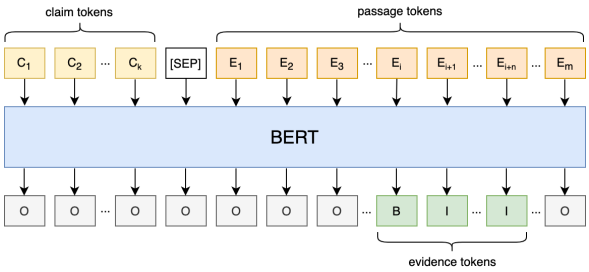

Figure 4: The sequence labeling model for evidence selection from a passage for a given claim.

$$L_\theta = -\sum_{i=1}^{N}\sum_{j=1}^{l_i} log(p_\theta(y_j^i)) \quad (3)$$

where N is the number of examples in the training batch, l_i the number of non-padding tokens of the i-th example and $p_\theta(y_j^i)$ the estimated softmax probability of the correct label for the j-th token of the i-th example. We trained the model on Factual-NLI with batch size 64, Adam optimizer and initial learning rate 5×10^{-5} until convergence.

3.5 Entailment Classification

We used an NLI model based on RoBERTa-large with a linear transformation of the [CLS] token embedding:

$$o = softmax(W \cdot \text{BERT}_{\text{[CLS]}}([P;H]) + b) \quad (4)$$

where P ; H is the concatenation of the premise with the hypothesis, $W_{3\times1024}$ a linear transformation matrix and $b_{3\times1}$ the added bias. We trained the model by minimizing the cross-entropy loss on the concatenation of three popular NLI datasets (SNLI, Multi-NLI and Adversarial-NLI) with batch size 64, Adam optimizer and initial learning rate 5×10^{-5} until convergence.

4 Results & Analysis

In this section we present the evaluation results of the three steps of the fact-checking process: retrieval, sentence selection and verification.

4.1 Passage Retrieval

For evaluation of passage retrieval, we report the recall@k and Mean Reciprocal Rank @100 on the

FEVER dataset, including the gold evidence passages and all the other passages from the June 2017 Wikipedia dump, and two extensions of FEVER, one with noise from web search results and another with additional synthetic examples from question answering datasets (Factual-NLI+). The results can be seen in Table 1. The dense retrieval model gives better ranking in the top 20 results compared to sparse retrieval, and the difference is significantly higher in the noisy FEVER dataset. We also evaluate a hybrid retrieval that merges results from the two models. The hybrid retrieval gives the highest recall across the three datasets. This shows that retrieval based on keyword matching is insufficient, and a dense encoder is required to surface many of the relevant passages from a noisy corpus.

FEVER						
Model	R@1	R@5	R@10	R@20	R@100	MRR@100
BM25	27.27	50.53	58.92	67.93	82.93	0.3807
Dense	43.81	65.47	69.61	72.51	75.71	0.5350
Hybrid	43.81	71.71	78.60	83.65	91.09	0.5557
FEVER + noise						
BM25	17.82	35.17	44.18	53.89	73.95	0.2649
Factual-NLI+						
BM25	25.23	45.02	53.20	61.56	77.96	0.3468
Dense	35.10	59.66	67.09	72.23	78.52	0.4608
Hybrid	34.91	61.29	70.03	77.51	87.90	0.4646

Table 1: Passage retrieval evaluation

4.2 Sentence Selection

For the evaluation of sentence selection, we calculate the token-level precision, recall and F1 score on Factual-NLI. We also examine how the various approaches perform on SciFact (Wadden et al., 2020), a domain-specific dataset for scientific claim verification. Table 2 shows the results. The proposed sequence labeling (SL) model gives the highest precision, recall and F1 score compared to the other two approaches. Especially the precision metric is significantly higher. Intuitively, the SL model has the best performance probably because it selects sentences based on their whole context. This mitigates the high false positive rate observed with the other two approaches: sentence-level entailment classification and embedding-based selection.

Training the SL model only using examples from FEVER, leads to worse performance, especially when evaluating the transferability to the scientific domain with the SciFact dataset (table 3). On SciFact, the sequence labelling model without further fine-tuning (line 2) comes second in terms of the F1 score after the sentence entailment selection model. However, after fine-tuning it on Sci-Fact, the SL model achieves the best F1 score. To conclude, the sequence labeling model seems to be the best solution for the selection of evidence sentences from passages compared to context agnostic approaches.

Model	P	R	F1
Sequence labeling (FEVER)	87.06	84.60	85.81
Sequence labeling (Factual-NLI)	94.78	92.11	93.43
Sentence Entailment	67.74	91.87	77.98
Embedding-based	66.12	90.29	76.34

Table 2: Sentence selection evaluation on Factual-NLI

Model	P	R	F1
Sequence labeling (FEVER)	48.30	43.48	45.77
Sequence labeling (Factual-NLI)	66.25	55.12	60.17
Sequence labeling (F-NLI + SciFact)	69.38	68.45	68.91
Sentence Entailment	62.21	71.54	66.55
Embedding-based	43.30	92.36	58.96

Table 3: Sentence selection evaluation on SciFact

4.3 Claim Verification

The claim verification is evaluated on the noisy extension of the Factual-NLI dataset. Firstly, we do the evaluation with oracle passage retrieval to examine how the various sentence selection approaches influence the verification accuracy. The results can be seen in table 5. An interesting observation is that when we pass the whole passages to the entailment classification model, the model tends to predict the neutral class most of the time. Selecting the correct evidence sentences alleviates this issue significantly (see figure 4). Finally, we do the evaluation for various combinations of passage retrieval and evidence selection approaches (table 4). Hybrid passage retrieval with results from BM25 and QR-BERT and sentence selection with the proposed sequence labeling model gives the best results.

5 The Quin System

The Quin system is based on the same three-stage architecture from our experiments; a hybrid passage retriever, a sentence selection model and an

Evidence	Supports			Refutes			Neutral			Macro Avg		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Oracle evidence	82.15	60.05	69.38	77.54	89.32	83.02	84.62	91.28	87.83	81.44	80.22	80.07
Whole passage	63.40	53.93	58.28	33.95	40.65	37.00	70.33	71.37	70.85	55.90	55.32	55.38
Sentence entailment	66.04	63.81	64.91	36.87	75.33	49.51	81.05	60.52	69.30	61.32	66.55	61.24
Embedding-based	72.92	69.96	71.41	57.37	74.64	64.87	83.30	78.33	80.74	71.20	74.31	72.34
Sequence labeling	74.40	56.68	64.34	75.27	81.96	78.47	81.67	88.32	84.87	77.11	75.66	75.89

Table 4: Claim verification evaluation on Factual-NLI with oracle passage retrieval

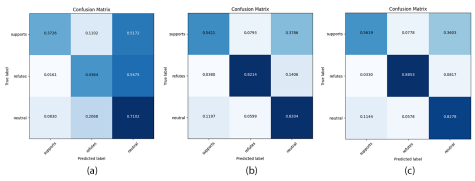

Figure 5: The confusion matrices for the entailment classification of the claim-passage pairs in Factual-NLI development set using: the whole passage (a), only evidence sentences from the sequence labeling model (b), and only golden evidence sentences (c).

Retrieval	Evidence	P	R	F1
Oracle	Oracle	81.44	80.22	80.07
BM25, k=100	Oracle			
Dense, k=100	Oracle			
Hybrid, k=100	Oracle			
Hybrid, k=100	Sentence entailment			
Hybrid, k=100	Embedding-based			
Hybrid, k=100	Sequence labeling			
Hybrid, k=20	Sequence labeling			
Hybrid, k=200	Sequence labeling			

Table 5: Claim verification on Factual-NLI+

NLI model. We deployed a demo[1] of the system for verifying claims about the COVID-19 pandemic from over 300K news articles and 100K research papers from the CORD-19 dataset (Wang et al., 2020). The system is able to retrieve relevant evidence for both statements and questions using a multi-task dense retriever trained on both Factual-NLI and MSMARCO. For statements, it classifies selected evidence in three categories (supporting, refuting, neutral), and based on the dominant class from multiple sources it returns a veracity rating (probably true, probably false, ambiguous and not enough evidence). For evidence selection from passages, we used the embedding-based selection method instead of the best-performing sequence labeling model. The reason is that this method seems to generalize much better in terms of recall, as seen from the evaluation on SciFact (table 3), and evidence recall is more important in a real-world fact-checking system, whose main utility is the discovery of relevant evidence rather than final judgements for claims. The importance of recall over precision can also be seen in table 4, where the embedding-based selection with significantly lower F1 score (due to much lower precision) leads to comparable claim verification performance with the sequence labeling model.

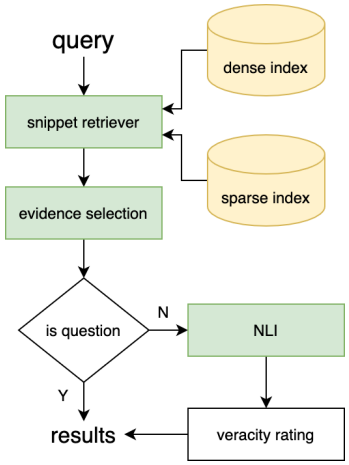

Figure 6: The architecture of the Quin system.

6 Conclusion & Future Work

In this paper we examined how the first stages of an automated fact-checking system can be improved. More specifically, we investigated how a dense passage retrieval model can lead to higher recall when retrieving relevant evidence, and proposed a sequence labeling model for selecting relevant evidence sentences for improving the final verification accuracy. For our experiments we also introduced an extension of the Factual-NLI dataset, that constitutes a more challenging benchmark for passage retrieval in fact-checking.

This work focused on improving the first two steps of a fact-checking system, without attempting to improve the last step, the entailment classi-

[1] https://quin.algoprog.com

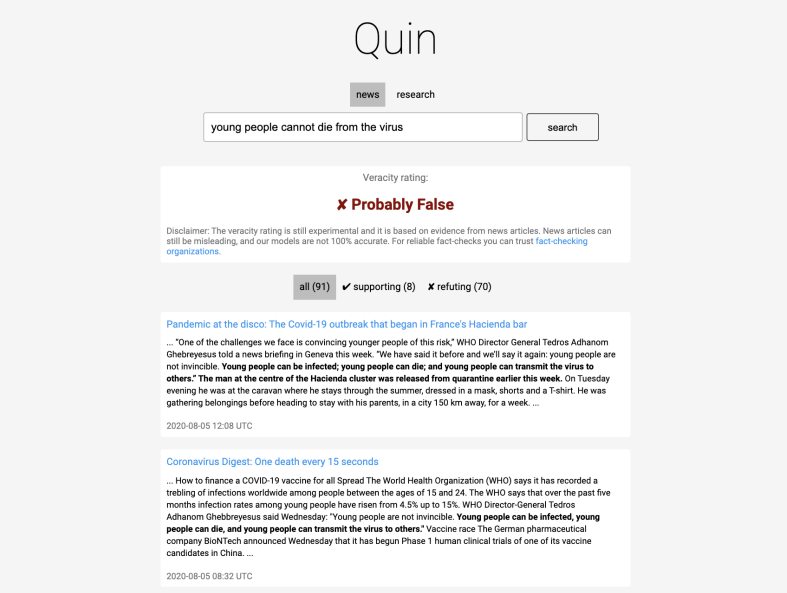

Figure 7: The Quin system returning relevant evidence and a veracity rating for a claim about COVID-19.

fication. Even though recent pre-trained language models seem to perform very well on the two popular NLI datasets (SNLI and Multi-NLI), they are not equally effective in a real-world setting. Recent work has identified bias in these two datasets (Poliak et al., 2018), which has a negative effect in the generalization ability of the trained models. The currently detected bias appears in the form of annotation artifacts in the hypothesis sentences. Because of this, models sometimes tend to ignore the information from the premise and assign a class based on the artifacts in the hypothesis. Various approaches have been proposed to unlearn or avoid this bias (He et al., 2019; Belinkov et al., 2019). Another weakness of these datasets, is that they are comprised of short, usually single-sentence, premises. As a consequence, models trained on these datasets usually do not perform well on noisy real-world data with multiple sentences. These issues led to the development of additional more challenging datasets, such as Adversarial NLI (Nie et al., 2020), which we also used for training our NLI model. Including examples from Adversarial-NLI improved the entailment classification F1 score from 74.14% to 80.07% (table 4), but still leaves a lot of room for improvement. Generalizable Natural Language Inference remains an open problem, and requires understanding of the reasons behind the misclassifications, the introduction of more diverse training examples without annotation artifacts and better language models.

References

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations.

Yonatan Belinkov, Adam Poliak, Stuart M. Shieber, Benjamin Van Durme, and Alexander M. Rush. 2019. Don't take the premise for granted: Mitigating artifacts in natural language inference. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 877–891. Association for Computational Linguistics.

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

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Wei-Cheng Chang, Felix X. Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training tasks for embedding-based large-scale retrieval. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

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.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. In *DeepLo@EMNLP-IJCNLP*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wentau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6086–6096. Association for Computational Linguistics.

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*.

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

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Rodrigo Nogueira, W. Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-stage document ranking with bert. *ArXiv*, abs/1910.14424.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Chris Samarinas, Wynne Hsu, and Mong Li Lee. 2020. Latent retrieval for large-scale fact-checking and question answering with nli training.

Tal Schuster, Darsh J Shah, Yun Jie Serene Yeo, Daniel Filizzola, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. *arXiv preprint arXiv:1908.05267*.

Dominik Stammbach and Elliott Ash. e-fever: Explanations and summaries for automated fact checking.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and verification. In *NAACL-HLT*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, L. Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

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*, abs/2004.14974.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, K. Funk, Rodney Michael Kinney, Ziyang Liu, W. Merrill, P. Mooney, D. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, B. Stilson, A. Wade, K. Wang, Christopher Wilhelm, Boya Xie, D. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. Cord-19: The covid-19 open research dataset. *ArXiv*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

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.