Improving Evidence Retrieval for Automated Explainable Fact-Checking

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

[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. In this paper, we describe the components of a three-stage automated fact-checking system, named Quin+. We demonstrate that using dense passage representations increases the evidence recall in a noisy setting. We experiment with two sentence selection approaches, an embedding-based selection using a dense retrieval model, and a sequence labeling approach for context-aware selection. Quin+ is able to verify open-domain claims using a large-scale corpus or web search results.

1 Introduction

With the emergence of social media and many individual news sources online, the spread of misinformation has become a major problem with potentially harmful social consequences. Fake news can manipulate public opinion, create conflicts, elicit unreasonable fear and suspicion. The vast amount of unverified online content led to the establishment of external post-hoc fact-checking organizations, such as PolitiFact, FactCheck.org, Snopes etc, with dedicated resources to verify claims online. However, manual fact-checking is time consuming and intractable on a large scale. The ability to automatically perform fact-checking is critical to minimize negative social impact.

Automated fact checking is a complex task involving evidence extraction followed by evidence reasoning and entailment. For the retrieval of relevant evidence from a corpus of documents, existing systems typically utilize traditional sparse retrieval which may have poor recall, especially when the relevant passages have few overlapping words with the claims to be verified. Dense retrieval models have proven effective in question answering as these models can better capture the latent semantic content of text. The work in (Samarinas et al., 2020) is the first to use dense retrieval for fact checking. The authors constructed a new dataset called Factual-NLI comprising of claim-evidence pairs from the FEVER dataset (Thorne et al., 2018) as well as synthetic examples generated from benchmark Question Answering datasets (Kwiatkowski et al., 2019; Nguyen et al., 2016). They demonstrated that using Factual-NLI to train a dense retriever can improve evidence retrieval significantly.

While the FEVER dataset has enabled the systematic evaluation of automated fact-checking systems, it does not reflect well the noisy nature of real-world data. Motivated by this, 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. We examine how dense representations can improve the first-stage retrieval recall of passages for fact-checking in a noisy setting, and make the retrieval of relevant evidence more tractable on a large scale.

However, the selection of relevant evidence sentences for accurate fact-checking and explainability remains a challenge. Figure 1 shows an example of a claim and the retrieved passage which has three sentences, of which only the last sentence provides the critical evidence to refute the claim. We propose two ways to select the relevant sentences, an embedding-based selection using a dense retrieval model, and a sequence labeling approach for context-aware selection. We show that the former generalizes better with a high recall, while the latter has higher precision, making them suitable for the identification of relevant evidence sentences. Our fact-checking system Quin+ is able

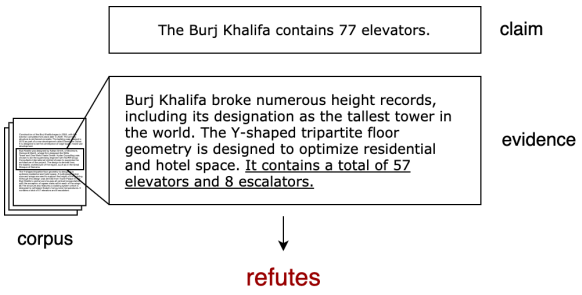

Figure 1: Sample claim and the retrieved evidence passage where only the last sentence is relevant.

to verify open-domain claims using a large corpus or web search results.

2 Related Work

Automated claim verification using a large corpus has not been studied systematically until the availability of the Fact Extraction and VERification dataset (FEVER) (Thorne et al., 2018). This dataset contains claims that are supported or refuted by specific evidence from Wikipedia articles. Prior to the work in (Samarinas et al., 2020), fact-checking solutions have relied on sparse passage retrieval, followed by a claim verification (entailment classification) model (Nie et al., 2019). Other approaches used the mentions of entities in a claim and/or basic entity linking to retrieve documents and a machine learning model such as logistic regression or an enhanced sequential inference model to decide whether an article most likely contains the evidence (Yoneda et al.; Chen et al., 2017; Hanselowski et al., 2018).

However, retrieval based on sparse representations and exact keyword matching can be rather restrictive for various queries. This restriction can be mitigated by dense representations using BERT-based language models (Devlin et al., 2019). The works in (Lee et al., 2019; Karpukhin et al., 2020; Xiong et al., 2020; Chang et al., 2020) have successfully used such models and its variants for passage retrieval in open-domain question answering. The results can be further improved using passage re-ranking with cross-attention BERT-based models (Nogueira et al., 2019). The work in (Samarinas et al., 2020) is the first to propose a dense model to retrieve passages for fact-checking.

Apart from passage retrieval, sentence (evidence) selection is also a critical task in fact-checking. These evidence sentences provide an explanation why a claim has been assessed to be credible or not. Recent works have proposed a BERT-based model for extracting relevant evidence sentences from multi-sentence passages (Atanasova et al., 2020). The authors observe that joint training on veracity prediction and explanation generation performs better than training separate models. The work in (Stammbach and Ash) investigates how the few-shot learning capabilities of the GPT-3 model (Brown et al., 2020) can be used for generating fact-checking explanations.

3 The Quin+ System

The automated claim verification task can be defined as follows: given a textual claim c and a corpus $D = \{d_1, d_2, ..., d_n\}$, where every passage d is comprised of sentences s_j, $1 \leq j \leq k$, a system will return a set of evidence sentences $\hat{S} \subset \bigcup d_i$ and a label $\hat{y} \in \{$probably true, probably false, inconclusive$\}$.

We have developed an automated fact-checking system, called Quin+, that verifies a given claim in three stages: passage retrieval from a corpus, sentence selection and entailment classification as shown in Figure 2. The label is determined as follows: we first perform entailment classification on the set of evidence sentences. When the number of retrieved evidence sentences that entail or contradict the claim is low, we label the claim as "inconclusive". Otherwise, if the number of evidence sentences that support the claim exceeds the number of sentences that refute the claim, we assign the label "probably true". In the opposite case, we assign the label "probably false".

3.1 Passage Retrieval

The passage retrieval model in Quin+ is based on a *dense* retrieval model called QR-BERT (Samarinas et al., 2020). This model 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 their dot product:

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

Dot product search can run efficiently using an approximate nearest neighbors index implemented using the FAISS library (Johnson et al., 2017). QR-BERT maximizes the sampled softmax loss:

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

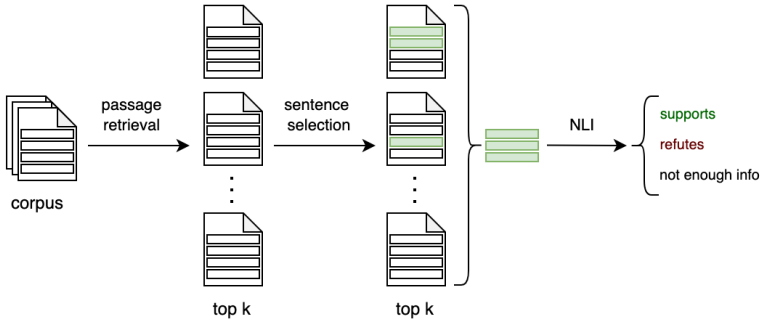

Figure 2: Three stages of claim verification in Quin+.

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 the batch b.

The work in (Samarinas et al., 2020) introduced the Factual-NLI dataset that extends the FEVER dataset (Thorne et al., 2018) with more diverse synthetic examples derived from question answering datasets. There are 359,190 new entailed claims with evidence and additional contradicted claims from a rule-based approach. To ensure robustness, we compile a new large-scale *noisy* version of Factual-NLI called Factual-NLI+[1]. This dataset includes all the 5 million Wikipedia passages in the FEVER dataset. We add 'noise' passages as follows. For every claim c in the FEVER dataset, we retrieve the top 30 web results from the Bing search engine and keep passages with the highest BM25 score that are classified as neutral by the entailment model. For claims generated from MSMARCO queries (Nguyen et al., 2016), we include the irrelevant passages that are found in the MSMARCO dataset for those queries. This results in 418,650 additional passages. The new dataset reflects better the nature of a large-scale corpus that would be used by real-world fact-checking system. We train a dense retrieval model using this extended dataset.

The Quin+ system utilizes a hybrid model that combines the results from the dense retrieval model described above and BM25 sparse retrieval to obtain the final list of retrieved passages. For efficient sparse retrieval, we used the Rust-based Tantivy full text search engine[2].

3.2 Sentence Selection

We propose and experiment with two sentence selection methods: an embedding-based selection and context-aware sentence selection method.

[1]https://archive.org/details/factual-nli

[2]https://github.com/tantivy-search/tantivy

The embedding-based selection method relies on the dense representations learned by the dense passage retrieval model QR-BERT. For a given claim c, we select the sentences from a given passage $d = \{s_1, s_2, ..., s_k\}$ whose relevance score $r(c, d)$ is greater than some threshold λ.

The context-aware sentence selection method uses a sequence labelling model. 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 (see Figure 3). We train a model based on RoBERTa-large (Liu et al., 2019), minimizing the cross-entropy loss:

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

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 $p_\theta(y_j^i)$ is the estimated softmax probability of the correct label for the j^{th} token of the i^{th} example. We train the sentence selection model in Quin+ on Factual-NLI with batch size 64, Adam optimizer and initial learning rate 5×10^{-5} until convergence.

3.3 Entailment Classification

Natural Language Inference (NLI), also known as textual entailment classification, is the task of detecting whether a hypothesis statement is entailed by a premise passage. It is essentially a text classification problem, where the input is a pair of premise-hypothesis (P, H) and the output a label $y \in \{$entailment, contradiction, neutral$\}$. An NLI model is often a core component of many automated fact-checking systems. Datasets like the Stanford Natural Language Inference corpus (SNLI) (Bowman et al., 2015), Multi-Genre

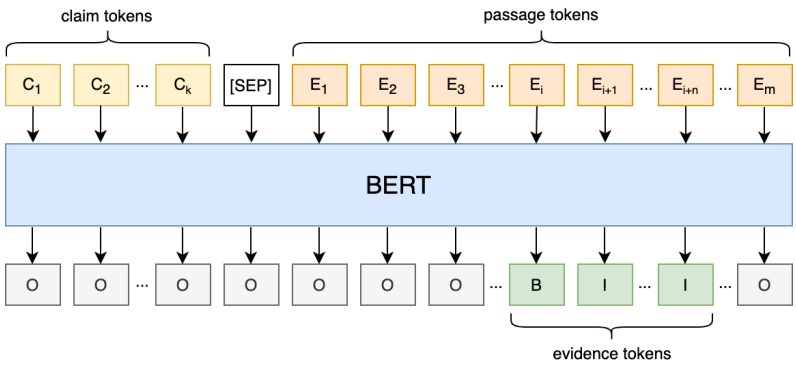

Figure 3: Sequence labeling model for evidence selection from a passage for a given claim.

Natural Language Inference corpus (Multi-NLI) (Williams et al., 2018) and Adversarial-NLI (Nie et al., 2020) have facilitated the development of models for this task.

Even though recent pre-trained NLI models seem to perform very well on the two popular NLI datasets (SNLI and Multi-NLI), they are not as 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. Another weakness of these datasets, is that they are comprised of short, usually single-sentence, premises. As a result, models trained on these datasets usually do not perform well on noisy real-world data involving multiple sentences. These issues have led to the development of additional more challenging datasets such as Adversarial NLI (Nie et al., 2020).

Our Quin+ system utilizes 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 train the entailment model by minimizing the cross-entropy loss on the concatenation of the 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 Performance of Quin+

We evaluate the three individual components of Quin+ (retrieval, sentence selection and entailment classification) and finally perform an end-to-end evaluation using various configurations.

Table 1 gives the recall@k and Mean Reciprocal Rank (MRR@100) of the passage retrieval models on FEVER and Factual-NLI+. We also compare the performance on a noisy extension of the FEVER dataset where additional passages from the Bing search engine are included as 'noise' passages. We see that when noise passages are added to the FEVER dataset, the gap between the hybrid passage retrieval model in Quin+ and sparse retrieval widens. This demonstrates the limitations of using sparse retrieval, and why it is crucial to have a dense retrieval model to surface relevant passages from a noisy corpus. Overall, the hybrid passage retrieval model in Quin+ gives the best performance compared to BM25 and the dense retrieval model.

(a) FEVER Dataset

Model	R@5	R@10	R@20	R@100	MRR
BM25	50.53	58.92	67.93	82.93	0.381
Dense	65.47	69.61	72.51	75.71	0.535
Hybrid	**71.71**	**78.60**	**83.65**	**91.09**	**0.556**

(b) FEVER with noise passages

Model	R@5	R@10	R@20	R@100	MRR
BM25	35.17	44.18	53.89	73.95	0.2649
Dense	54.10	62.13	68.09	75.24	0.4053
Hybrid	**54.89**	**64.61**	**73.33**	**86.11**	**0.4074**

(c) Factual-NLI+ Dataset

Model	R@5	R@10	R@20	R@100	MRR
BM25	45.02	53.20	61.56	77.96	0.347
Dense	59.66	67.09	72.23	78.52	0.461
Hybrid	**61.29**	**70.03**	**77.51**	**87.90**	**0.465**

Table 1: Performance of passage retrieval models.

Table 2 shows the token-level precision, recall and F1 score of the proposed sentence selection methods on the Factual-NLI dataset and a domain-specific (medical) claim verification dataset, SciFact (Wadden et al., 2020). We also compare the

(a) Factual-NLI Dataset

Model	Precision	Recall	F1
Baseline	67.74	91.87	77.98
Sequence labeling	**94.78**	**92.11**	**93.43**
Embedding-based	66.12	90.29	76.34

(b) SciFact Dataset

Model	Precision	Recall	F1
Baseline	62.21	71.54	66.55
Sequence labeling	**69.38**	68.45	**68.91**
Embedding-based	43.30	**92.36**	58.96

Table 2: Performance of sentence selection methods.

performance to a baseline sentence-level NLI approach, where we perform entailment classification (using the model described in Section 3.3) on each sentence of a passage and select the non-neutral sentences as evidence. We observe that the sequence labeling model gives the highest precision, recall and F1 score when tested on the Factual-NLI dataset. Further, the precision is significantly higher than the other methods.

On the other hand, for the SciFact dataset, we see that sequence labeling method remains the top performer in terms of precision and F1 score after fine-tuning, although its recall is lower than the embedding-based method. This shows that sequence labeling model is able to mitigate the high false positive rate observed with the embedding-based selection method by taking into account the surrounding context.

The Factual-NLI+ dataset contains claims with passages that either support or refute the claims with some sentences highlighted as ground truth specific evidence. Table 3 shows the performance of the entailment model to classify the input evidence as supporting or refuting the claims. The input evidence can be in the form of the whole passage, ground truth evidence sentences, or sentences selected by our sequence labeling model. We observe that the entailment classification model performs poorly when whole passages are passed as input evidence. However, when the specific sentences are passed as input, the precision, recall, and F1 measures improve. The reason is that our entailment classification model is trained mostly on short premises. As a result, it does better on sentence-level evidence compared to the longer passages.

Finally, we carry out an end-to-end evaluation of our fact-checking system on Factual-NLI+ us-

(a) Supporting evidence

Input	Precision	Recall	F1
Whole passages	63.40	53.93	58.28
Highlighted ground truth	82.15	60.05	69.38
Selected sentences	74.40	56.68	64.34

(b) Refuting evidence

Input	Precision	Recall	F1
Whole passages	33.95	40.65	37.00
Highlighted ground truth	77.54	89.32	83.02
Selected sentences	75.27	81.96	78.47

Table 3: Performance of entailment classification model on different forms of input evidence.

Passage retrieval	Sentence selection	F1
BM25, k=5	Embedding-based	52.76
BM25, k=20	Embedding-based	47.65
BM25, k=5	Sequence labeling	49.65
Dense, k=5	Embedding-based	49.03
Dense, k=5	Sequence labeling	52.83
Dense, k=50	**Sequence labeling**	**58.22**
Hybrid, k=6	Embedding-based	50.29
Hybrid, k=6	**Sequence labeling**	**57.24**
Hybrid, k=50	Sequence labeling	52.60

Table 4: End-to-end claim verification on Factual-NLI+ for different configurations.

ing various configurations of top-k passage retrieval (BM25, dense, hybrid, for various values of $k \in [5, 100]$) and evidence selection approaches (embdedding-based and sequence labeling). Table 4 shows the macro-average F1 score for the three classes (supporting, refuting, neutral) for some of the tested configurations. We see that dense or hybrid retrieval with evidence selection using the proposed sequence labeling model gives the best results. Even though hybrid retrieval seems to lead to slightly worse performance, it requires much less passages (6 instead of 50) and makes the system more efficient.

5 System Demonstration

We have created a demo for verifying open-domain claims using the top 20 results from a web search engine. For a given claim, Quin+ returns relevant text passages with highlighted sentences. The passages are grouped into two sets, supporting and refuting. It computes a veracity rating based on the number of supporting and refuting evidence. It returns "probably true" if there are more supporting evidence, otherwise it returns "probably false". When the number of retrieved evidence is low, it returns "inconclusive". Figure 4 shows a screen dump of the system with a claim

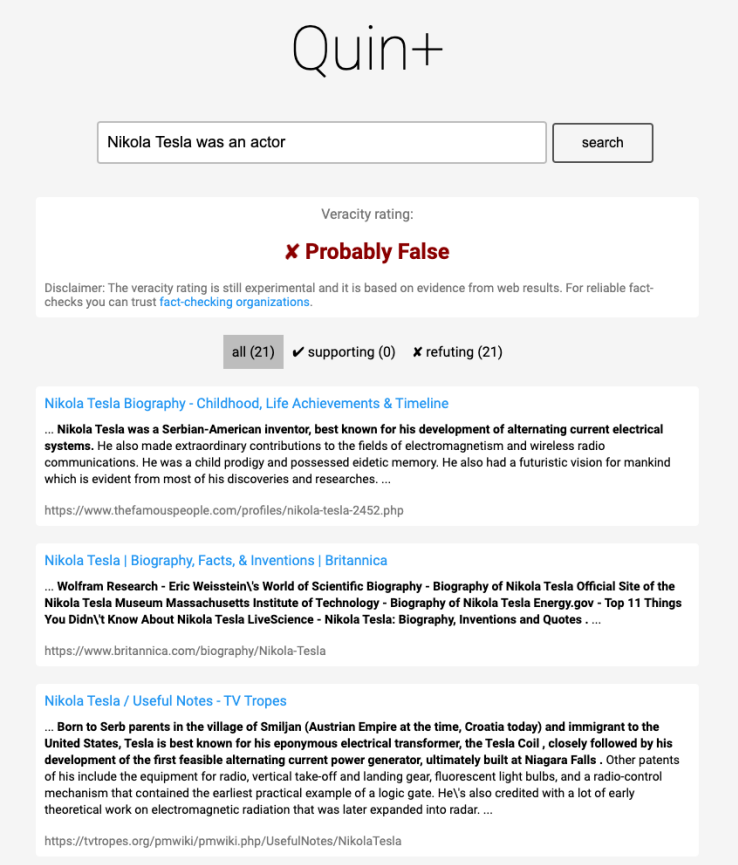

Figure 4: The Quin+ system returning relevant evidence and a veracity rating for a claim.

that has been assessed to be probably false based on the overwhelming number of refuting sentence evidence (21 refute versus 0 support). Quin+ can also be used on a large-scale corpus.

6 Conclusion & Future Work

In this work, we have presented the components of our three-stage fact-checking system Quin+. We have demonstrated how a dense retrieval model can lead to higher recall when retrieving passages for fact-checking. In addition, we have proposed two schemes to select relevant sentences: an embedding-based approach and a sequence labeling model to improve the claim verification accuracy. Quin+ gave promising results in our extended Factual-NLI+ corpus, and is also able to verify open-domain claims using web search results. The source code of our system is publicly available[3].

For the future development of large-scale fact-checking systems we believe that a new benchmark needs to be introduced. The currently available datasets, including Factual-NLI+, are not suitable for evaluating the performance of systems that have a large number of evidence sources such as news publishers or research papers.

[3]https://github.com/algoprog/Quin

References

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

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.

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 Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

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*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268.

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. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 941–948. IEEE.

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*.

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.

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. 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.