Constrained Fact Verification for FEVER

Adithya Pratapa*, **Sai Muralidhar Jayanthi***, **Kavya Nerella***
Language Technologies Institute
Carnegie Mellon University
{vpratapa,sjayanth,knerella}@cs.cmu.edu

Abstract

Fact-verification systems are well explored in the NLP literature with growing attention owing to shared tasks like FEVER. Though the task requires reasoning on extracted evidence to verify a claim's factuality, there is little work on understanding the reasoning process. In this work, we propose a new methodology for fact-verification, specifically FEVER, that enforces a closed-world reliance on extracted evidence. We present an extensive evaluation of state-of-the-art verification models under these constraints.

1 Introduction

A rapid increase in the spread of misinformation on the Internet has necessitated automated solutions to determine the validity of a given piece of information. To this end, the Fact Extraction and VERification (FEVER) shared task (Thorne et al., 2018a)[1] introduced a dataset for *evidence-based fact verification*. Given a claim, the task involves extracting relevant evidence sentences from a given Wikipedia dump and assigning a label to the claim by reasoning over the extracted evidence (SUPPORTS / REFUTES / NOTENOUGHINFO).

Several recent works (Liu et al., 2020; Soleimani et al., 2020; Zhao et al., 2020) leverage representations from large pre-trained language models (LMs) like BERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2019) to achieve state-of-the-art results on FEVER. However, it is unclear how factual knowledge encompassed in these LMs influences the verification process.

More recently, Lee et al. (2020) developed a fact verification system solely based on large pre-trained LMs and presented their superior zero-shot performance on FEVER compared to a random baseline. This result clearly shows the influence of factual knowledge embedded inside these LMs, but relying entirely on such knowledge directly contrasts to the *evidence-based* paradigm of fact-verification. Such reliance can be problematic, especially with evolving evidence (Wikipedia pages are constantly updated to reflect the latest events). Schuster et al. (2019) illustrate this phenomenon through an example fact, "*Halep failed to ever win a Wimbledon title*", which was valid until July 2019 but not thereafter.

In this work, we propose methods to train fact-verification models that explicitly reason on the available evidence instead of relying on the factual knowledge in pre-trained LMs, thereby emulating a closed-world setting. This is particularly important in the context of the FEVER dataset because of the overlap between the source corpus used for compiling FEVER and the ones commonly used to pre-train LMs (*Wikipedia*).

We build upon the work of Clark et al. (2020) that demonstrated the ability of transformers (BERT, RoBERTa) to function as soft theorem provers. They induce a closed-world reasoning process by fine-tuning on a carefully curated synthetic natural language rulebase. In this work, we transfer this ability to FEVER and gauge the feasibility of such closed-world reasoning. Additionally, we also construct an entity-anonymized version of FEVER following Hermann et al. (2015) for evaluating our proposed models. We construct the anonymized version by masking prominent named entities in the claim-evidence pairs, thereby reducing any reliance on pre-trained factual knowledge.

Our experiments adopt the popular three-stage pipeline of FEVER task, comprising document selection, evidence sentence extraction, and claim verification (Thorne et al., 2018b). We primarily focus on the claim verification stage of FEVER, while using the state-of-the-art document selec-

* equal contribution
[1] https://fever.ai/

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 7826–7832,
November 16–20, 2020.

tion and evidence sentences extraction from Liu et al. (2020). Our focus is motivated since only the claim verification step involves a joint (often complicated) reasoning over the extracted evidence. Our main contributions are,

- We propose various pre-training strategies for large pre-trained LMs to induce a closed-world setting during fact verification in FEVER.
- We adapt an existing synthetic natural language rulebase to FEVER by incorporating NOTENOUGHINFO label.
- We create an anonymized version of the FEVER dataset to facilitate investigation into the factual knowledge through named entities.

Our datasets and code are publicly available.[2]

2 Constrained Verification

Traditionally, most FEVER systems rely on large pre-trained language models (LMs) to encode the claim and extracted evidence sentences. Previously, Schuster et al. (2019) studied various reasons for the surprisingly good performance of claim-only classifiers on FEVER and reported dataset idiosyncrasies to be the primary reason as opposed to world knowledge in word embeddings. However, they present only a preliminary analysis of the impact of world knowledge from GloVe embeddings (Pennington et al., 2014). In this work, we present an in-depth analysis because the issue is particularly relevant in the context of large pre-trained LMs. To the best of our knowledge, we are not aware of any other works that look into the impact of embedding's world knowledge on FEVER.

In a nutshell, we model the task under a closed-world setting with the extracted evidence as the only available factual information to the model. Overall, we believe the methods proposed in this paper are general enough to apply to any fact-verification task. However, we show a case study only on FEVER due to its wide-spread popularity.

To this end, we first present an entity-anonymized version of the FEVER dataset and then propose pre-training strategies to enforce the above described closed-world setting on FEVER models.

2.1 Anonymization

A straightforward way to discourage the use of prior factual knowledge in fact-verification systems is to anonymize the named entities. An intuitive way to achieve this is to replace them with a custom list of abstract entity markers. We adapt a related technique from reading comprehension literature (Hermann et al., 2015) to our task. Given a pair of claim and extracted evidence sentences, we first identify the set of named entities from Wiki-titles of evidence sentences. We then replace all the occurrences of these named entities with abstract markers sampled randomly from a predefined list. We present an anonymized FEVER instance in Table 1. We use the resulting anonymized FEVER dataset to evaluate our proposed methods.

Claim	
David Carradine$_{\text{ent0}}$ was not in Kung Fu$_{\text{ent1}}$	
Evidences	
David Carradine$_{\text{ent0}}$	He was known for his leading role as a peace loving Shaolin monk, Kwai Chang Caine, in the 1970s television series Kung Fu$_{\text{ent1}}$
Kung Fu (TV series)$_{\text{ent1}}$	Kung Fu$_{\text{ent1}}$ is an American action adventure martial arts western drama television series starring David Carradine$_{\text{ent0}}$

Table 1: Example from Anonymized FEVER dataset. Each evidence constitutes the Wiki-title and a corresponding sentence. The two named entities (ent0, ent1) are highlighted.

2.2 Towards Closed-World FEVER

Clark et al. (2020) analyze the logical reasoning capabilities of transformer-based models on a variety of question-answering and reading comprehension tasks. Given a question and a context comprising of a set of simple facts and rules in natural language, models are expected to reason only based on the provided context, thereby emulating the ability to perform closed-world reasoning. They propose a synthetic training dataset (henceforth referred to as RuleTaker dataset) to fine-tune pre-trained models like RoBERTa. They observe high performances ($\geq$95% accuracy) on the synthetic test set, motivating us to adapt a similar training methodology for FEVER.

Table 2 shows an example context from the RuleTaker dataset. Each question-context pair in this dataset belongs to one of the following types, **Type–A**: provable/disprovable statements, can be labeled by reasoning directly over the context, **Type–B**: unprovable statements, reasoning

[2] https://github.com/adithya7/constrained-fever

Facts/Triples		
F_1: Bob is blue.		
F_2: Fiona is kind.		
Rules		
R_1: All white people are red.		
R_2: Blue people are white.		
R_3: If someone is red then they are kind.		
RuleTaker-CWA: Questions		
Question	**Proof**	**Our Label**
$Q1$. Fiona is kind.	(F_2)	SUPPORTS
$Q2$. Bob is white.	($F_1 \rightarrow R_2$))	SUPPORTS
$Q3$. Bob is not red.	(($F_1 \rightarrow R_2$) $\rightarrow R_1$)	REFUTES
$Q4$. Fiona is red.	CWA	NOTENOUGHINFO
$Q5$. Bob is not round.	CWA	NOTENOUGHINFO
RuleTaker-Skip–fact: Questions		
Question	**New Context**	**Our Label**
$Q1$. Fiona is kind.	[F_1, F_2, R_1, R_2, R_3]	SUPPORTS
$Q2$. Bob is white.	[F_1, F_2, R_1, R_2, R_3]	SUPPORTS
$Q3$. Bob is not red.	[F_1, F_2, R_1, R_2, R_3]	REFUTES
$Q6$. Fiona is kind.	[F_1, ~~F_2~~, R_1, R_2, R_3]	NOTENOUGHINFO
$Q7$. Bob is white.	[~~F_1~~, F_2, R_1, R_2, R_3]	NOTENOUGHINFO
$Q9$. Bob is not red.	[~~F_1~~, F_2, R_1, R_2, R_3]	NOTENOUGHINFO

Table 2: Example from the compiled RuleTaker-CWA and RuleTaker-Skip–fact.

over the context is not sufficient to conclude these statements.[3]

The RuleTaker dataset assigns a TRUE or FALSE label to each question-context pair. Type–A were labeled by reasoning over the context, whereas Type–B were labeled by invoking the closed-world assumption (CWA) ($Q4$, $Q5$ in Table 2). The provided context (facts and rules) constitutes the closed-world setup. Moreover, Type–A are additionally annotated with a proof constituting a reasoning chain over a subset of facts and rules.

We adapt the RuleTaker dataset to FEVER by introducing a new NOTENOUGHINFO label for unprovable question-context pairs. In particular, we construct two FEVER-style RuleTaker datasets, namely RuleTaker-CWA and RuleTaker-Skip–Fact (example in Table 2).

RuleTaker-CWA: We convert all the labels for Type–B pairs into NOTENOUGHINFO ($Q4$, $Q5$ in Table 2) and relabel TRUE and FALSE from Type–A into SUPPORTS and REFUTES respectively ($Q1$, $Q2$, $Q3$ in Table 2).

RuleTaker-Skip–Fact: For each Type–A question, we create a contrastive setting by removing a necessary fact (i.e., required in proof) from the original context. The label for the modified question-context pair becomes NOTENOUGHINFO because the question can no longer be answered under the modified context ($Q6$, $Q7$, $Q8$ in Table 2). We also retain the original Type–A pairs by converting all TRUE and FALSE labels to SUPPORTS and REFUTES respectively ($Q1$, $Q2$, $Q3$ in Table 2). To maintain a balanced dataset, we randomly sample a fraction of newly created NOTENOUGHINFO labels. Note that we only work with Type–A pairs in this variant. Occasionally there could be multiple valid proofs for the same question-context pair. We currently ignore these questions to avoid inconsistencies arising from other valid reasoning methods over the modified context. Table 3 presents the statistics for the train, dev and test splits in the proposed RuleTaker-CWA and RuleTaker-Skip–fact datasets.

As a natural adaptation, we also considered creating a similar Skip–fact variant of the FEVER dataset. Each claim in FEVER was annotated with potentially many evidence sets, and each evidence set can consist of multiple evidence sentences. Ideally, we need all sentences within single evidence set to validate the claim, i.e., it requires multi-hop reasoning. Unfortunately, we noticed cases where a proper subset of an evidence set is enough to prove/disprove the claim (see Table 4).

Split/Label	SUPPORTS	REFUTES	NOTENOUGHINFO
RuleTaker-CWA			
Train	32034	32034	55000
Validation	4581	4581	7832
Test	9156	9156	15680
RuleTaker-Skip–fact			
Train	27360	27369	25389
Validation	3891	3906	3647
Test	7720	7733	7165

Table 3: Distribution of compiled RuleTaker datasets.

2.3 Methodology

We now present the methodology to train constrained fact-verification models for the FEVER shared task. Many state-of-the-art FEVER models use the standard BERT encoder (Devlin et al., 2019) to encode a concatenation of claim and evidence sentences. To enforce closed-world reasoning over available evidence, we first pre-train the BERT encoder on the proposed variants of Rule-

[3]Type–A, Type–B correspond to the proof types {proof, inv-proof}, and {rconc, inv-rconc, random, inv-random} respectively, from the original dataset.

Claim
Roman Atwood is a content creator.
(One) Gold Evidence Set
1. (Roman Atwood) Roman Bernard Atwood (born May 28, 1983) is an American YouTube personality, comedian, vlogger and pranker.
2. (Comedian) A popular saying, variously quoted but generally attributed to Ed Wynn, is, "A comic says funny things; a comedian says things funny", which draws a distinction between how much of the comedy can be attributed to verbal content and how much to acting and persona.

Table 4: An example from the FEVER dataset. Wikipedia page titles for the evidence sentences are mentioned in parentheses. Even though the original dataset contains both evidence sentences within a single evidence set, we can label the given claim using just the first evidence sentence. Such cases would result in erroneous labels when creating Skip–fact version of FEVER.

Model	# labels	Test accuracy
RuleTaker	2	90.5
RuleTaker-CWA	3	92.5
RuleTaker-Skip–fact	3	91.4

Table 5: RuleTaker results on individual in-domain test sets. Note, these are separate test sets.

Taker datasets following Clark et al. (2020).

Firstly, the reasoning models in Clark et al. (2020) were first trained on the RACE multi-choice question answering dataset (Lai et al., 2017) and then fine-tuned on the RuleTaker dataset. In our experiments, we follow the same pipeline (including hyper-parameters) except to replace original RuleTaker dataset with our adaptations, RuleTaker-CWA and RuleTaker-Skip–fact.[4] In Table 5, we present the results of the pretrained RuleTaker-CWA and RuleTaker-Skip–fact on their respective test sets. In general, we notice high performance on the synthetic test sets, indicating the model's ability to rely only on available evidence.

We now utilize the above fine-tuned BERT encoders (CWA, Skip–fact) with two state-of-the-art graph-based reasoning networks for claim verification, KGAT (Liu et al., 2020) and Transformer-XH (Zhao et al., 2020), as well as a robust BERT-based classifier.

BERT–concat: Evidence sentences retrieved before claim verification are concatenated to the claim along with their Wiki-titles and are encoded using a pretrained BERT encoder. The [CLS] representation from the encoder is then directly used for classification.[5]

KGAT (Liu et al., 2020): A kernel-based graph attention network over the evidence graph. Each node in the graph encodes a concatenation of individual evidence sentence (along with Wiki-title) and the claim. Knowledge propagation between the nodes of this graph is achieved using a Gaussian *edge kernel* on a word-word similarity matrix, while individual node importance is measured using a separate *node kernel*. The initial node representations are refined using the above kernels and a single graph attention layer.

Transformer-XH (Zhao et al., 2020): Evidence graph is constructed and initialized in a way similar to KGAT, but the knowledge propagation between the nodes is achieved using special *eXtra-Hop attention* mechanism. For each node, the [CLS] token embedding from BERT is considered as an *attention hub* and is revised using a combination of the extra-hop attention and the traditional *in-sequence* attention.[6]

We compare the above-proposed curricula (CWA, Skip–fact) against a baseline curriculum (Original) where we initialize the verification models with standard pretrained BERT weights (bert-base-cased). We use huggingface transformers (Wolf et al., 2019) in all of our experiments.[7]

3 Experiments

For each of the three models, BERT–concat, Transformer-XH, and KGAT, we show results on the three different training curricula, Original, CWA, and Skip–fact in Table 6. We evaluate all our trained models on three datasets, the official dev-set of FEVER task (Std.), symmetric FEVER v0.2

[4]Note that we use the depth-3ext-NatLang set, which constitutes *depth*=3 dataset augmented with 10% each of *depth*=0,1,2 and crowdsourced natural language. We refer to the original work for more details about the RuleTaker dataset construction process.

[5]We use BertAdam optimizer with learning rate 3e-5, train for ten epochs and choose the best checkpoint based on dev label accuracy

[6]For KGAT and Transformer-XH, we follow the same hyper-parameters as the original work. We refer the readers to the original papers for more details.

[7]https://huggingface.co/transformers/

Eval set	BERT–concat			Transformer-XH			KGAT		
	Original	CWA	Skip–fact	Original	CWA	Skip–fact	Original	CWA	Skip–fact
Std.	**77.3**	74.8	74.3	76.7	74.6	74.5	**77.5**	73.8	73.6
Symm.	**57.5**	51.6	55.1	55.4	51.0	**59.0**	28.0	17.1	14.4
Anon.	73.2	68.0	70.9	70.4	68.1	65.8	**74.3**	70.8	69.1

Table 6: Label Accuracy on Standard (Std.), Symmetric (Symm.) and Anonymized (Anon.) *dev* sets. We highlight the best results in each row (evaluation set).

Eval set	BERT–concat (Anon. train)		
	Original	CWA	Skip–fact
Anon.	**75.7**	73.8	73.6

Table 7: Performance of BERT–concat model trained on anonymized FEVER train dataset. We report the accuracies on anonymized *dev* set.

(Schuster et al., 2019) (Symm.), and our proposed anonymized version of Std. (Anon.). Symmetric FEVER proposed by Schuster et al. (2019) constructs three adversarial claim-evidence pairs based on the original pair from the FEVER dev set.

On most evaluation sets, we found the models trained with Original curriculum performed better than our proposed curricula (CWA, Skip–fact) except on symmetric FEVER where Transformer-XH with Skip–fact does slightly better. Across the models, we notice a considerable drop in performance on Anon. set, validating our hypothesis about existing reliance on factual knowledge. To see the individual impact of the entity-anonymization, we train the BERT–concat model on train split of Anon. FEVER dataset. We observe improvements across the three curricula, with Original still outperforming the proposed curricula (Table 7).

Through our constrained verification setup, we expect the models to reason using only the extracted evidence. The evidence retrieval from Liu et al. (2020) achieves a recall of 94%, indicating the feasibility of reasoning only on extracted evidence in FEVER. With Original outperforming the proposed strategies on both the standard and anonymized FEVER, we find that world knowledge is helpful for FEVER.

Limitations Firstly, our anonymization is a regex-based method and relies only on the entities in Wiki-titles, and this might be insufficient for handling ambiguous titles. Secondly, the Rule-Taker dataset's domain is significantly different from that of the FEVER dataset, thereby presenting a challenge in re-using the pretrained encoder. Additionally, it is not entirely clear as to what constitutes the world (or factual) knowledge for a given task and as highlighted by Clark et al. (2020), effectively combining implicit pretrained knowledge (from encoders) with explicitly stated knowledge (from evidence) remains a challenge.

4 Related Work

We adopted the widely used document selection method from Hanselowski et al. (2018). Many recent state-of-the-art FEVER systems involve reasoning over evidence graphs (Zhou et al., 2019; Zhong et al., 2019; Liu et al., 2020; Zhao et al., 2020) along with competitive LM-based models (Soleimani et al., 2020). Dataset specific idiosyncrasies have been identified in FEVER (Thorne et al., 2019; Schuster et al., 2019) as well as in NLI (Gururangan et al., 2018; Poliak et al., 2018; Naik et al., 2018; McCoy et al., 2019), but is not the focus of this work.

5 Conclusion

We identify a critical issue with existing claim verification systems, especially the recent models that utilize large pre-trained LMs. We propose to perform fact verification under a closed-world setting and present our results on the task of FEVER. While it is hard to evaluate the reliance on implicit pretrained knowledge, our initial results indicate that such reliance is helpful for FEVER.

Acknowledgments

We would like to thank Aditi Chaudhary and Graham Neubig for their insightful discussions on this work. We also thank the anonymous reviewers for their valuable feedback.

References

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-PRICAI-20*. International Joint Conferences on Artificial Intelligence Organization.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1693–1701, Cambridge, MA, USA. MIT Press.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Nayeon Lee, Belinda Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabsa. 2020. Language models as fact checkers? In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 36–41, Online. 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*.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351, Online. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

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.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3419–3425, Hong Kong, China. Association for Computational Linguistics.

Amir Soleimani, Christof Monz, and Marcel Worring. 2020. Bert for evidence retrieval and claim verification. In *Advances in Information Retrieval*, pages 359–366, Cham. Springer International Publishing.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. 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, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2019.

Evaluating adversarial attacks against multiple fact verification systems. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2944–2953, Hong Kong, China. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Chen Zhao, Chenyan Xiong, Corby Rosset, Xia Song, Paul Bennett, and Saurabh Tiwary. 2020. Transformer-xh: Multi-evidence reasoning with extra hop attention. In *International Conference on Learning Representations*.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2019. Reasoning over semantic-level graph for fact checking. *arXiv preprint arXiv:1909.03745*.

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 the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901, Florence, Italy. Association for Computational Linguistics.