The Second Fact Extraction and VERification (FEVER2.0) Shared Task

James Thorne
University of Cambridge
jt719@cam.ac.uk

Andreas Vlachos
University of Cambridge
av308@cam.ac.uk

Oana Cocarascu
Imperial College London
oana.cocarascu11@imperial.ac.uk

Christos Christodoulopoulos
Amazon
chrchrs@amazon.co.uk

Arpit Mittal
Amazon
mitarpit@amazon.co.uk

Abstract

We present the results of the second Fact Extraction and VERification (FEVER2.0) Shared Task. The task challenged participants to both build systems to verify factoid claims using evidence retrieved from Wikipedia and to generate adversarial attacks against other participant's systems. The shared task had three phases: *building, breaking and fixing*. There were 8 systems in the builder's round, three of which were new qualifying submissions for this shared task, and 5 adversaries generated instances designed to induce classification errors and one builder submitted a fixed system which had higher FEVER score and resilience than their first submission. All but one newly submitted systems attained FEVER scores higher than the best performing system from the first shared task and under adversarial evaluation, all systems exhibited losses in FEVER score. There was a great variety in adversarial attack types as well as the techniques used to generate the attacks, In this paper, we present the results of the shared task and a summary of the systems, highlighting commonalities and innovations among participating systems.

1 Introduction

Significant progress for a large number of natural language processing tasks has been made through the development of new deep neural models. Higher scores for shared tasks such as Natural Language Inference (Bowman et al., 2015) and Question Answering (Rajpurkar et al., 2016) have been achieved through models which are becoming increasingly complex. This complexity raises new challenges: as models become more complex, it becomes difficult to fully understand and characterize their behaviour. From an NLP perspective, there has been an ongoing discussion as to what extent these models understand language (Jia and Liang, 2017) or to what extent they are exploiting unintentional biases and cues that are present in the datasets they are trained on (Poliak et al., 2018; Gururangan et al., 2018). When a model is evaluated on data outside of the distribution defined (implicitly) by its training dataset, its behaviour is likely to be unpredictable; such "blind spots" can be exposed through *adversarial evaluation* (Szegedy et al., 2014).

The first Fact Extraction and VERification (FEVER) shared task (Thorne et al., 2018b) focused on building systems that predict whether a textual claim is SUPPORTED or REFUTED given evidence (see (Thorne et al., 2018a) for a task description), or NOTENOUGHINFORMATION in case Wikipedia does not have appropriate evidence to verify it. As automated systems for fact checking have potentially sensitive applications it is important to study the vulnerabilities of these systems, as well as the deficiencies of the datasets they are trained on. Such vulnerabilities were also the motivation behind Ettinger et al. (2017)'s NLP shared task that was inspired by the Build It, Break It, Fix It competition[1].

The second Fact Extraction and VERification (FEVER2.0) shared task is building on the dataset of the first FEVER shared task, but adopted the setup of build-it, break-it, fix-it where *builders* submitted systems based on the original FEVER dataset and task definition; *breakers* generated adversarial examples targeting the systems built in the first stage; and finally, *fixers* implemented solutions to remedy the attacks from the second

[1] https://builditbreakit.org

Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER), pages 1–6
Hong Kong, November 3, 2019. ©2019 Association for Computational Linguistics

stage.

In this paper, we present a short description of the task and dataset, present a summary of the submissions and the leader board, and highlight future research directions.

2 Task Description

2.1 Task Phases

In what follows we describe the three phases of FEVER2.0 in more detail:

Build-It In the first phase of the shared task, "builders" constructed fact verification systems that were trained using the FEVER dataset released in Thorne et al. (2018a). Participants were required to submit docker images of systems which implemented a common web API that would facilitate interactive development of attacks through a sandbox which was hosted for the duration of the shared task.

The top 4 submission from the first shared task were submitted as baseline systems for this shared task: UNC (Nie et al., 2019), UCLMR (Yoneda et al., 2018), Athene (Hanselowski et al., 2018) and Papelo (Malon, 2018).

Break-It In the second phase, "breakers", were tasked with generating adversarial examples that induce classification errors for the existing systems. Breakers submitted a dataset of up to 1000 instances with equal number of instances for each of the three classes (SUPPORT, REFUTE and NOTENOUGHINFORMATION); half of which were released to fixers and half of which were retained as a blind test set. We considered only novel claims (i.e. not contained in the original FEVER dataset) as valid entries to the shared task. All of the claims in this submission were annotated were annotated by the shared task organizers for quality assurance and to measure correctness.

To aid with preparing their submission of 1000 instances, the organizers hosted a web-based sandbox. Breakers had access to 8 systems (4 top systems from the first FEVER shared task (Thorne et al., 2018b), the baseline from (Thorne et al., 2018a) and 3 new qualifying submissions from the 'Build-It' phase) that were hosted by the shared task organisers. Participants could experiment with attacks by submitting small samples of 50 instances for scoring twice a day via a shared task portal which returned FEVER scores of all the hosted systems.

Fix-It Using the adversarial examples, the original builders or teams of dedicated "fixers" incorporate the data generated from the "break-it" phase to improve the system classification performance and resilience to adversarial attack.

2.2 Scoring Method

The submissions were scored using 'potency' and 'resilience' (Thorne et al., 2019) that compute a weighted average of FEVER scores: accounting for the correctness of adversarial instances.

Potency Intuitively, better adversarial instances induce more classification errors, resulting in a lower FEVER score of the systems they are evaluated on. We measure the effectiveness of breakers' adversarial instances (a) on a builder's system (s) through the average reduction in FEVER score (from a perfect system) on the set of predictions made by the system $\hat{Y}_{s,a}$. The score is weighted by the correctness rate c_a of the adversarial instances. Instances are correct if they are grammatical, appropriately labeled and meet the annotation guidelines requirements described by Thorne et al. (2018a).

$$\text{Potency}(a) \overset{\text{def}}{=} c_a \frac{1}{|S|} \sum_{s \in S} \big(1 - f(\hat{Y}_{s,a}, Y_a)\big)$$

Resilience A system that is resilient will have fewer errors induced by the adversarial instances, reflected in higher scores at evaluation. We wish to penalize systems for making mistakes on instances from adversaries with higher correctness rate. We define *resilience* of a system s as the weighted average FEVER score, weighted by the correctness rate for each adversary, $a \in A$:

$$\text{Resilience}(s) \overset{\text{def}}{=} \frac{\sum_{a \in A} c_a \times f(\hat{Y}_{s,a}, Y_a)}{\sum_{a \in A} c_a}$$

For the 'build-it' phase, we report both FEVER score of the system over the FEVER shared task test set (Thorne et al., 2018a) and the resilience of the system over the FEVER2.0 test set that comprises adversarial instances submitted by the breakers. For the 'break-it' phase, we report the potency of attack over all systems and the correctness rate. For the 'fix-it' phase, we report the score delta compared to the system submitted in the 'build-it' phase.

3 Participants and Results

System	Resilience (%)	FEVER Score (%)
Papelo	37.31	57.36
UCLMR	35.83	62.52
DOMLIN	35.82	68.46
CUNLP	32.92	67.08
UNC	30.47	64.21
Athene	25.35	61.58
GPLSI	19.63	58.07
Baseline	11.06	27.45

Table 1: Results from the FEVER2.0 Builder phase. Italicised systems are from the original FEVER shared task – submitted as reference systems for FEVER2.0.

System	Correct Rate (%)	Potency (%)
TMLab	84.81	66.83
CUNLP	81.44	55.79
NbAuzDrLqg	64.71	51.54
Rule-based Baseline	82.33	49.68
Papelo*	91.00	64.79

Table 2: Results from the FEVER2.0 Breaker phase. *Papelo's submission contained only NOTENOUGHINFO claims which did not qualify for the shared task. Its potency is reported, but is not included in the calculations for resilience of the systems.

System	FEVER Score (%)	Resilience (%)
CUNLP	68.80 (+1.72)	36.61 (+3.69)

Table 3: Results from the FEVER2.0 Fixer phase.

3.1 Builders Phase

Team DOMLIN (Stammbach and Neumann, 2019) used the document retrieval module of Hanselowski et al. (2018) and a BERT model for two-staged sentence selection based on the work by (Nie et al., 2019). They also use a BERT-based model for the NLI stage.

The CUNLP team (Hidey et al.) used a combination of Google search and TF-IDF for document retrieval and a pointer network using features from BERT and trained with reinforcement learning.

Finally, team GPLSI (Alonso-Reina et al., 2019) kept Hanselowski et al. (2018)'s document retrieval and NLI modules. For the sentence selection they converted both the claims and candidate evidence sentence into OpenIE-style triples using the extractor from Estevez-Velarde et al. (2018) and compared their semantic similarity.

3.2 Breakers Phase

The TMLab (Niewinski et al., 2019) adversarial claims were generated with Generative Enhanced Model (GEM). GEM is a modified and fine-tuned GPT-2 language model fed with text sampled from two hyperlinked Wikipedia pages and additional keyword input. Claims were labeled by annotators and the evidence sentences were manually added. In addition, the team manually generated claims with SUPPORTS labels to ensure class balance in their submission.

One of the shortcomings of the original FEVER dataset was the lack of complex claims that would require multi-hop inference or temporal reasoning and the CUNLP team designed their adversarial attacks along these principles (Hidey et al.). They produce multi-hop reasoning claims by augmenting existing claims with conjunctions or relative clauses sourced from linked Wikipedia articles. For temporal reasoning adversarial examples they use hand-written rules to manipulate claims containing dates, for example changing "in 2001" to "4 years before 2005" or "between 1999 and 2003". Finally, they create noisy versions of existing claims by using entities that have a disambiguation page in Wikipedia and by using the lexical substitution method of Alzantot et al. (2018).

Team NbAuzDrLqg (Kim and Allan, 2019) submitted mostly manually created adversarial claims targeting the retrieval as well as the NLI components of FEVER systems. For the retrieval attacks, the team created claims that didn't contain enti-

Breaker	Attack	FEVER Score (%)	Label Accuracy (%)	*n*
CUNLP	Multi-Hop Reasoning	31.54 ± 13.19	51.64 ± 7.18	130
	Multi-Hop Temporal Reasoning	8.33 ± 2.08	24.48 ± 16.98	24
	Date Manipulation	27.53 ± 6.07	34.18 ± 4.50	94
	Word Replacement	28.87 ± 6.79	29.08 ± 9.28	71
	Conjunction	38.25 ± 18.01	42.50 ± 15.93	50
	Phrasal Additions	55.63 ± 13.16	55.63 ± 20.22	20
NbAuzDrLqg	NotEnoughInfo	76.39 ± 34.33	76.39 ± 34.33	18
	SubsetNum	0.00 ± 0.00	16.12 ± 17.08	38
TMLab	AI Generated	38.07 ± 13.29	40.63 ± 11.04	44
	Paraphrase	0.00 ± 0.0	43.06 ± 19.59	9

Table 4: Breakdown of attack type for each breaker and average FEVER scores and Label accuracy for the 8 systems used in the shared task. n = total number of instances of this class

ties that could be used as query terms for evidence documents/sentences. To target the NLI component, the team created attacks based on arithmetic operations, logical inconsistencies, and vague or hedged statement. Some of these attack types failed to meet the guidelines of the shared task and were not marked as correct instances by annotators: these have been excluded from the analysis in Section 4, Table 4.

Finally, team Papelo submitted only NOTENOUGHINFO claims and therefore did not meet the requirements of submitting a balanced dataset. While the potency results for this method are reported, it does not qualify for the shared task and this attack is not used in computation of system resilience.

The rule-based baseline system is a version of the adversary described in Thorne et al. (2019) where string transformations are applied to claims to generate new instances. The rules were manually constructed regular expression patterns that match common patterns of claims in the dataset and perform both label-altering and label-preserving changes.

3.3 Fixers Phase

The only submission to this phase was from the CUNLP team (Hidey et al.). Based on their own attacks during the Breakers phase they sought to make improvements in multi-hop retrieval and temporal reasoning. To improve multi-hop retrieval, they introduce an additional document pointer network trained with the top 4 layers of a fine-tuned BERT Wikipedia title-to-document classifier as input features. They also improve sentence selection by modeling the sequence of relations at each time step through training a network to predict a sequence pointers to sentences in the evidence. For temporal reasoning they employ a set of arithmetic rules on top of predicate arguments extracted with an OpenIE system. As seen in Table 3 they improve their system's FEVER score, but more importantly they increase its resilience by 3.69%.

4 Analysis

In the 'break-it' phase of the competition, breakers submitted adversarial instances that were designed to induce classification errors in fact verification systems. The shared task solicited metadata with each instance that described how the attack was generated. In Table 4 we report the FEVER score and accuracy of the systems for each of the breaker's attack types. We report only instances that were annotated as 'correct' and attack types with more than 5 instances.

There were two attack types which had a FEVER score of 0: the Paraphrase attack from TMLab and the SubsetNum attack from NbAuzDrLqg. While some systems returned the correct label, no system had the combination of the correct label and evidence. The Multi-Hop and Multi-Hop Temporal Reasoning attacks from CUNLP also induced a high number of errors in the systems.

The SubsetNum attack from NbAuzDrLqg was a template-based attack which required transitive reasoning with respect to the area and size of geographic regions. The Multi-Hop claims from CUNLP were manually generated to require inference that combines evidence from multiple entities. Both these types of attacks highlight limitations of systems when performing inductive reasoning and composition of knowledge.

The TMLab paraphrase attack strategy was to re-write sentences from Wikipedia articles in terms borrowed from different texts (not included in evidence set) to mislead the systems. This highlighted a limitation of all systems as while correct labels were being applied, correct evidence was not identified in any of these cases. This attack had a higher potency than TMLab's other automated submission, 'AI Generated', which generated claim text from the Generative Enhanced Model (GEM). Similar to CUNLP, correctly classifying these claims requires compositional knowledge and reasoning with information from multiple Wikipedia pages.

5 Conclusions

The second Fact Extraction and VERification shared task received three qualifying submissions for the builder round and three qualifying submissions for the breaker round and one fixer submission. All of the breakers submitted adversarial instances that were more potent than the rule-based baseline presented in Thorne et al. (2019). In this paper we summarized the approaches, identifying commonalities and features that could be further explored.

Future work will continue to address limitations in human-annotated evidence and explore other ways in which systems can be made more robust in predicting the veracity of information extracted from real-world untrusted sources.

Acknowledgements

The authors with to thank the participants of the first FEVER shared task for making their code and models available and for being responsive to queries. The authors with to thank Tim Baldwin and Trevor Cohn for their advice.

References

Aimée Alonso-Reina, Robiert Sepúlveda-Torres, Estela Saquete, and Manuel Palomar. 2019. Team GPLSI. approach for automated fact checking. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. 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.

S Estevez-Velarde, Y Gutierrez, A Montoyo, A Piad-Morffis, R Munoz, and Y Almeida-Cruz. 2018. Gathering object interactions as semantic knowledge. In *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, pages 363–369.

Allyson Ettinger, Sudha Rao, Hal Daumé III, and Emily M Bender. 2017. Towards linguistically generalizable NLP systems: A workshop and shared task. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 1–10.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation Artifacts in Natural Language Inference Data.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. 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.

Tuhin Hidey, Christopherand Chakrabarty, Tariq Alhindi, Siddharth Varia, Kriste Krstovski, Mona Diab, and Smaranda Muresan. Non archival shared task submission.

Robin Jia and Percy Liang. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems.

Youngwoo Kim and James Allan. 2019. FEVER breaker's run of team NbAuzDrLqg. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics.

Christopher Malon. 2018. Team Papelo: Transformer networks at FEVER. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 109–113, Brussels, Belgium. Association for Computational Linguistics.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Piotr Niewinski, Maria Pszona, and Maria Janicka. 2019. GEM: Generative enhanced model for adversarial attacks. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis Only Baselines in Natural Language Inference. (1):180–191.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. pages 2383–2392.

Dominik Stammbach and Guenter Neumann. 2019. Team DOMLIN: Exploiting evidence enhancement for the fever shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *International Conference on Learning Representations*.

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

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2019. Evaluating adversarial attacks against multiple fact verification systems. In *EMNLP-IJCNLP*.

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.

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