FEVER Breaker's Run of Team NbAuzDrLqg

Youngwoo Kim and **James Allan**
Center for Intelligent Information Retrieval
University of Massachusetts Amherst
Amherst, MA 01003
{youngwookim, allan}@cs.umass.edu

Abstract

We describe our submission for the Breaker phase of the second Fact Extraction and VERification (FEVER) Shared Task. Our adversarial data can be explained by two perspectives. First, we aimed at testing model's ability to retrieve evidence, when appropriate query terms could not be easily generated from the claim. Second, we test model's ability to precisely understand the implications of the texts, which we expect to be rare in FEVER 1.0 dataset. Overall, we suggested six types of adversarial attacks. The evaluation on the submitted systems showed that the systems were only able get both the evidence and label correct in 20% of the data. We also demonstrate our adversarial run analysis in the data development process.

1 Introduction

The Fact Extraction and Verification (FEVER) workshop focuses on developing fact-check systems, which can resolve "fake-news" and misinformation problems. In the shared task of FEVER, the goal is to develop a system which can verify the given claim, by retrieving evidence from the documents from Wikipedia and classifying the claim into either *Supports*, *Refutes*or *NotEnoughInfo*. While the systems in the shared task of first FEVER workshop (FEVER 1.0) showed impressive performance, it was questionable if they are robust against adversarial claims that are different from the test data from the original dataset (Thorne and Vlachos, 2019).

The second workshop on Fact Extraction and VERification has a shared task that can investigate the robustness of the systems. The shared task is in a Build it Break it Fix it setting. In the first phase, participants (Builders) develop fact-check systems as what was done in last year's shared task. In the second phase, participants (Breakers) will have access to the systems and *attack* the systems to generate claims which are challenging for the builders. In the third phase, the Fixers would fix the systems to be robust toward the Breakers' claims.

We participated in the second phase (Breakers Run) in the competition. We submitted 203 instances over seven types of attacks. For 6 out of 7 attack types (except SubsetNum), the claims were manually written. The claims for SubsetNum were generated based on a template.

Our submission resulted in Raw Potency of 79.66% but resulted in bad Correct Rate of 64.71% and the Adjusted Potency of 51.54%.

Our data were annotated to have 25.7% as incorrect label and 22.8% as ungrammatical, which includes 8.9% overlap. While the ungrammatical cases evenly appeared among all the cases, incorrect label cases are concentrated in NotClear attack.

We consider there are two types of challenges for the Fact-Checking system. The first is retrieval challenge and the second is language understanding challenge.

The results of the Fever 1.0 showed that the most of the evidences can be found among the candidate sentences that are retrieved by taking the terms in the claim as a query (Yoneda et al., 2018; Hanselowski et al., 2018a).

Three of our attacks focuses on retrieval challenges. The claims from EntityLess attack have few entities that can be used to retrieve evidence documents. The claims from EntityLinking different name from that which is in the evidence sentence, so the system need to link other name from the other article that explains alternative names for an entity. The claims from SubsetNum require 3 sentences as the evidence, where two of the evidence document can be found from the terms of the claim, but the other evidence cannot.

Remaining three attacks focuses on precise understanding of the text. We considered the case that the relevant article mentions the claim, but another sentence from the article says the claim to be not true (Controversy) or to be not clear (NotClear). If the system blindly picks most relevant sentences, the system can miss such clarifying information. The claims from FiniteSet consider the cases that some expression can imply that no more event of the particular type can happen other than the mentioned events.

In section 2, we explain our motivations for the attack types. In section 3, we explain how we generate 6 types of attacks. In section 4 we discuss the shared task results. In addition to actual submission results, section 5 discuss about the analysis in adversarial attack development phase

2 Design motivations

The claims of original FEVER dataset are made from the randomly chosen sentences (Thorne et al., 2018). We expect that many sentences share similar semantic patterns, while there are only few sentences that have different pattern than the majority. Randomly sampling sentences would result in many claims that can be handled by similar fact checking strategies, which makes the dataset hard to contain challenging and exceptional claims that are less trivial to fact-check. Here is an example of exceptional claims. Given a sentence, if the claim is entailed by the sentence, it is okay to conclude *Supports* for most cases. However, there are a few cases that the following sentence denies what's written in the previous sentence. Our attack types Controversy and NotClear test such cases.

In relation extraction domain, it was considered as a serious challenge to have ability to disambiguate a polysemous entity mention or infer that two orthographically different mentions are the same entity (Rao et al., 2013). We refer this challenge as entity linking and suggest that entity linking should be more intensely tested for fact-checking task. In the FEVER 1.0, many system solely relied on the neural network to handle entity linking. For the names of entities that are mentioned often in the corpus, word embedding could be trained enough to handle it. We expect that neural network might fail when it comes to the rarely mentioned surface names. We expect that original FEVER dataset will not have many such cases.

	Supports	Refutes	NE	Total
EntityLess	1	7	2	10
EntityLinking	8	1	0	9
SubsetNum	50	50	0	100
Controversy	0	10	0	10
NotClear	0	0	34	34
FiniteSet	4	6	0	10
NE	0	0	30	30
Total	63	74	66	203

Table 1: Label statistics for our submission. NE stands for *NotEnoughInfo*.

3 Claim generation for each type of attacks.

Our submission includes six types of adversarial cases and one type that only contain *NotEnoughInfo*to make all of three labels to have similar number of claims. Examples for the six attacks are listed in Table 2 and Table 3.

3.1 EntityLess[1]

This attack contains case that the evidence articles cannot be easily searched by the words in the claim. The claims only contains more common terms such as 'university', 'alumni' and 'U.S.'. In the example in the Table 2, the evidence is in 'Harvard University' article, while the important term 'Harvard' is not given in the claim. We expect that the system would wrongly answer *NotEnoughInfo*.

3.2 EntityLinking

This case tests the ability to identify different surface names for the same entity. The collection has the sentences that introduce multiple names of an entity. We selected one of such sentences which we expect to be not too popular and it is used as a first evidence. As a second evidence, we searched the sentence that mentions the entity and replaced the name of the entity with another name. We expect that the system would wrongly answer *NotEnoughInfo*.

3.3 SubsetNum

This case is generated based on a simple logic: if region A is part of B and B is smaller than C, A is smaller than C. In the example is Table 2, the sec-

[1]This attack was originally named 'TwoHops' in our submission.

No	Type	Claim	Label
1	EntityLess	No university has more than 5 alumni who became U.S. presidents.	Refutes
2	EntityLinking	Kanha Tiger Reserve has a significant population of swamp deer.	Supports
3	SubsetNum	The area of Nerva, Spain is larger than the area of Madhya Pradesh.	Refutes
4	Controversy	September Dossier revealed the fact that Iraq had reconstituted its nuclear weapons programme.	Refutes
5	NotClear	In 1899 Arnold Droz-Farny proved Droz-Farny line theorem.	NE
6	FiniteSet	Since 1960, no person was executed for his crime in Republic of Ireland.	Refutes

Table 2: Claims and the each cases of attack described in section 3. NE is for NotEnoughInfo

No	Evidence
1	[Harvard University] Harvard's alumni include eight U.S. presidents,
2	[Kanha Tiger Reserve] The park has a significant population of Bengal tiger, Indian leopards, the sloth bear, **barasingha** and Indian wild dog. (...) The **barasingha**, also called **swamp deer**, (....)
3	[Province of Huelva] Its area is **10,148 km^2.** [Nerva, Spai] Nerva is a town and municipality located in the province of Huelva, southern Spain. [Madhya Pradesh] Its total area is **308,252 km^2.**
4	[September Dossier] The dossier even **alleged** that Iraq had reconstituted its nuclear weapons programme. Without exception, all of the allegations included within the September Dossier have been since proven to be **false**, as shown by the Iraq Survey Group.
5	[Droz-Farny line theorem] The theorem was stated by Arnold Droz-Farny in 1899, but it is **not clear** whether he had a proof.
6	[Michael Manning (murderer)] Michael Manning was an Irish murderer who became the twenty-ninth and **last** person to be executed in the Republic of Ireland. The execution by hanging was duly carried out on 20 April **1954** (...)

Table 3: Evidences for the claims of Table 2. The words in bracket are the title of the article. Evidence 5 is not actually an evidence because the label is NotEnoughInfo. The sentence was listed to show that it might be mistakenly considered as an evidence.

	OK	GR	UN	UN,GR	Total	Correct Rate
EntityLess	2	1	2	0	5	0.60
EntityLinking	3	1	1	0	5	0.57
SubsetNum	36	7	3	1	47	0.40
Controversy	4	2	1	0	7	0.20
NotClear	3	0	9	8	20	0.15
FiniteSet	1	3	1	0	5	0.77
NE	12	0	0	0	12	1.00

Table 4: Acceptability judgments.
- OK : The claim is grammatical and the label is supported by the evidence.
- GR : The claim is ungrammatical.
- UN : The claim is grammatical but the label is incorrect.

ond and third evidence could be directly retrieved from the claim, but not the first evidence.

The claims were automatically generated. We extracted the information using the predefined templates. We first extracted the list of the entities that refer to regions. Then we extracted subset relations. The area information of each entity was parsed. We expect that the system would wrongly answer *NotEnoughInfo*.

3.4 Controversy

This case tests if the system can distinguish the mentions that are not actually true. Two evidence sentences are required. A sentence suggests information and the following sentence says that the previous statement is not true. All the claims for these cases are *Refutes*. We expect that the system would wrongly answer *Supports*.

3.5 NotClear

Wikipedia has sentences that say "It is not clear ..." (Table 2). We wrote the claims that are mentioned to be not clear and we consider this implies *NotEnoughInfo*. Because the label is *NotEnoughInfo*, we did not include the evidences.

The annotators did not accepted most of the claims (85%) and annotated they are not *NotEnoughInfo*(including the one in the table). It is not clear if they accepted the sentences with 'not clear' as evidences or they found from other documents. We expect that the system would wrongly answer *Supports*or *Refutes*.

3.6 FiniteSet

A sentence "A is ninth and last to do B." implies that there are only nine possible events for B. Moreover, if another event is claimed to be happened at the time which is later than when A happened, it cannot be true. For many cases keyword 'last' is just enough to restrict the times. Both *Supports* and *Refutes* cases are generated. We expect that the system would wrongly answer *NotEnoughInfo*.

3.7 NE

Our adversarial claims are mostly *Supports* or *Refutes*. In order to make each label has same similar number of claims we add claims whose label is *NotEnoughInfo*. These claims are not particularly adversarial compared to others.

4 Task Evaluation

The breaker's runs were evaluated by the following metrics:

$$\text{Potency(b)} \stackrel{\text{def}}{=} \frac{1}{|S|}\sum_{s\in S}(1 - \text{FEVER}(\mathcal{Y}_{s,b})) \quad (1)$$

$$\text{Adjusted\_Potency}(b) \stackrel{\text{def}}{=} r_{accept} \times Potency(b)$$

FEVER($\mathcal{Y}_{s,b}$) is the official evaluation metric, which is roughly the fraction of the instances that got both the evidences and label correct.

Our submission resulted in the raw potency of 79.66%. Accepted rate was 64.71%. Adjusted potency was 51.54%.

The raw potency of 79.66 implies that systems only got 20% got correct. Considering that 15% of the whole data was NE category which was not actually adversarial, the systems totally fail on our adversarial data.

During the shared task, we tested each type of attack on the running docker images of the shared task test server

For the final Fixer phase, the accepted instances from all breaker's run were collected. The collected data were provided to the fixers so that the systems can be revised or re-trained on the adversarial data. There was only one fixer system (CUNLP) and it showed FEVER score of 32.92% before they fixed the system. After they fixed the system it achieved the FEVER score of 68.80%. Note that these scores for the fixer system are results of all breaker's submissions not only our submission.

We were not provided the performance for the only our runs, but still we can make some speculation about the potency of adversarial instances in this shared task. We expect that the adversarial runs were rather limited in their diversity, the fixer was able to revise this challenges either manually or by machine learning models ability to adapt to new types of data.

5 Development Analysis

Here, we show a few adversarial instances that we generated during the development process. Note that some of these claims (2, 4) are of different categories from what was introduced in section 3, because they were not included in final submission. We evaluated these claims on the provided

No	Attack Type	Claim	Label
1	Time	Barack Obama is the first USA president to be born in America.	Refutes
2	SubsetSum	Idonesia does have the larger population than the town of Abu Al-Khaseeb	Supports
3	EntityLess	More than 10 people have walked on the moon.	Supports
4	Numeric	Borneo is larger than Crete Island	Supports
5	Controversy	Apollo astronauts did not actually walk on the Moon.	Refutes

Table 5: Claims tested in our development phase

No	Evidence
1	[Bill Clinton] William Jefferson Clinton (born William Jefferson Blythe III; August 19, 1946) is an American politician who served as the **42nd** president of the United States from 1993 to 2001. (...) Clinton was born and raised in **Arkansas** (...) [Barack Obama] **Barack** Hussein **Obama** II (born August 4, 1961) is an American attorney and politician who served as the **44th** president of the United States from 2009 to 2017. [Arkansas] Arkansas is a state in the southern region of the **United States**
2	[Indonesia] With over 261 million people, it is the world's 4th most populous country (...) [Abu Al-Khaseeb] Abu Al-Khaseeb (sometimes spelled Abu Al-Khasib) is a town in Abu Al-Khaseeb District, Basra Governorate, southern Iraq. [Iraq] Around 95% of the country's 37 million citizens are Muslims, with Christianity, Yarsan, Yezidism and Mandeanism also present
3	[List of Apollo astronauts] Twelve of these astronauts walked on the Moon 's surface or [Harrison Schmitt] (...) he also became the **twelfth** and second-youngest person to set foot on the Moon.
4	[Borneo] Borneo is the third-largest island in the world and the largest in Asia [Crete] Crete is the largest and most populous of the Greek islands, the 88th largest island in the world
5	[List of Apollo astronauts] Twelve of these astronauts walked on the Moon 's surface

Table 6: Evidences for the claims of Table 2.

sandbox interface, which runs the previously submitted systems. The systems are UCL (Yoneda et al., 2018), Athens (Hanselowski et al., 2018b), UCL-MR (Yoneda et al., 2018), Papelo (Malon, 2018), GPLSI, Columbia and the baseline system (Thorne et al., 2018).

The claims and evidences are listed in Table 5 and 6. Claim 1 in the Table 5 requires fact-check system to collect and combine many evidences. The system has to check if there are presidents who were born in America and precede Barack Obama's term. Claim 2 is an example of the previously explained SubsetSum attack. Claim 3 could be challenging because it does not contain any good keyword in it. It also requires systems to be able to compare numbers. Claim 4 requires to compare numbers. We expected systems could make mistake as evidence sentences have numerous "largest" in them. Claim 5 has related documents that could be mistakenly taken as evidence to support the claim. There is an article "Moon landing conspiracy theories", which contains sentence saying "12 Apollo astronauts did not actually walk on the Moon". Because this evidence sentence is very similar to the claim in terms of term matching, this might be retrieved as an evidence and might confuse the system.

Table 7 shows the results of each systems, mainly focusing on if the systems get the classification labels correct. The systems rarely select the evidences that we submitted. However, as there are many alternative evidences for these claims, we could conclude this as total failure.

	UNC	Athene	UCL MR	Papelo	GPLSI	Columbia	baseline
1	X	O	O	X	X	X	O
2	O	X	X	X	O	X	O
3	X	O	X	X	O	X	O
4	**O**	X	X	X	O	X	X
5	X	O	O	X	X	X	O

Table 7: Results of each system on the claims of Table 5. O and X denote if the system correctly get the classification label. Only one case had both the label and evidences correct: UNC on the 4th claim. Claim 1, 3, 4 and 5 are underlined to denote that they may have many possible evidences.

6 Conclusion

This year's FEVER shared task showed that currently systems for fact checking are sensitive to these adversarial attacks. To develop robust systems for fact check, we need to build better evaluation dataset which contains challenging and diverse test instances.

Acknowledgement

The authors wish to thank anonymous reviewers for their helpful advice and the FEVER Organizers for their efforts in managing the insightful workshop.

References

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018a. Ukp-athene: Multi-sentence textual entailment for claim verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018b. Ukp-athene: Multi-sentence textual entailment for claim verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108.

Christopher Malon. 2018. Team papelo: Transformer networks at fever. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 109–113.

Delip Rao, Paul McNamee, and Mark Dredze. 2013. Entity linking: Finding extracted entities in a knowledge base. In *Multi-source, multilingual information extraction and summarization*, pages 93–115. Springer.

James Thorne and Andreas Vlachos. 2019. Adversarial attacks against fact extraction and verification. *arXiv preprint arXiv:1903.05543*.

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

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.