

# FEVER: a large scale dataset for Fact Extraction and VERification

James Thorne  
University of Sheffield  
j.thorne@sheffield.ac.uk

Andreas Vlachos  
University of Sheffield  
a.vlachos@sheffield.ac.uk

Christos Christodoulopoulos  
Amazon Research Cambridge  
chrchrs@amazon.co.uk

Arpit Mittal  
Amazon Research Cambridge  
mitarpit@amazon.co.uk



The University  
Of  
Sheffield.



## Motivation

- Accurately extracting information from text documents is essential for natural language technologies.
- How can we verify if the information is correct by checking it against encyclopedic articles?

## This Work

- We introduce a new dataset containing 185,000 true and false facts written by human annotators.
- For each claim, we:
  - add evidence from multiple Wikipedia pages at a sentence level
  - label supported/refuted/not enough info given the evidence.
- Both evidence and label must be correct for scoring.** This leads us towards building accountable systems, where a justification/explanation of the verdict is provided.

## Dataset Construction

- Claim generation:** sentence sampled from intro sections of 50,000 most popular Wikipedia pages. Annotator writes simple sentences for each fact in the original sentence.

World knowledge can be introduced in controlled manner from a dictionary (using hyperlinked pages on Wikipedia)
- Claim Mutation:** for each claim, annotator makes 6 modifications akin to relations in Natural Logic Inference (negation, generalization, specialization, substitution etc.)
- Claim Labelling:** different annotator selects a set of sentences that completely support or refute a given claim. Evidence can be combined from multiple pages.

Claim	The <b>Rodney King Riots</b> took place <b>in the most populous county</b> in the <b>USA</b>
Evidence	[wiki/Los Angeles Riots]: The 1992 Los Angeles riots, <b>also known as the Rodney King</b> riots were a series of riots, lootings, arsons, and civil disturbances that <b>occurred in Los Angeles County</b> , California in April and May 1992.  [wiki/Los Angeles County]: <b>Los Angeles County</b> , officially the County of Los Angeles, <b>is the most populous county</b> in the <b>United States</b> .
Verdict	Supported

Generating Claims About

Warren Buffett

Source Sentence

This is the sentence that is used to substantiate your claims about Warren Buffett

Buffett has been the chairman and largest shareholder of Berkshire Hathaway since 1970, and his business exploits have had him referred to as the "Wizard", "Oracle" or "Sage" of Omaha by global media outlets.  
[Show Context](#)

Dictionary

Click the word for a definition. These definitions can be used to support the claims you write or make the claims more complex by making a deduction using the dictionary definitions  
  
The dictionary comes from the blue links on Wikipedia. This may be empty if the passage from Wikipedia contains no links.

[Berkshire Hathaway](#)  
Berkshire Hathaway Inc. is an American multinational conglomerate holding company headquartered in Omaha, Nebraska, United States.  
[List of assets owned by Berkshire Hathaway](#)  
[Omaha, Nebraska](#)  
Omaha is the largest city in the state of Nebraska and the county seat of Douglas County.  
[shareholder](#)

True Claims (one per line)

Aim to spend about 2 minutes generating 2-5 claims from this source sentence

Warren Buffett is the chairman of an American multinational company.  
Warren Buffet's company is based in the largest city in the state of Nebraska

## Quality Assurance and Human Evaluation

- Information Retrieval:** How annotators with time constraints against *super-annotators* with no time restrictions? Precision: 95.42%. Recall: 72.36%
- Inference:** Are the annotators reaching the same verdict with the evidence they find? We sampled 4% of claims and compute 5-way IAA; Kappa: 0.6841 (n=7506)
- Human Evaluation:** We (authors) re-annotated 227 claims, found 91.2% annotated correctly.
- Lessons Learned:** Hard to remove annotator's world knowledge. Hard to come up with 'universal' definitions.

## Catch the FEVER:

PARTICIPATE IN THE SHARED TASK  
SUBMIT A PAPER TO THE WORKSHOP

Check the FEVER website for more details:  
<http://fever.ai>



## Baseline Evaluation

We provide baselines using simple and state of the art methods for information retrieval and textual entailment:

- Evidence Retrieval - DrQA** (Chen et al., 2017)  
Trade-off number of documents/sentences (recall) against pipeline RTE accuracy. **Upper-bound Score: 62.8%**
- Recognizing Textual Entailment**
  - Multilayer Perceptron (MLP)** (Riedel et al., 2017)
  - Decomposable Attention (DA)** (Parikh et al., 2016)

	MLP	DA	DA (SNLI)
Accuracy (%)	65.13	80.82	38.54

- Full Pipeline**

Model	Accuracy (%)	FEVER Score (%)
MLP	41.86	19.04
DA	52.09	32.57

- Future Areas to Explore:**  
Multi-sentence natural language inference  
(baseline model concatenates the sentence strings)

Trade-off between volume of evidence (Recall) and the accuracy of the downstream inference component

Negative sampling strategy for training textual entailment classifier has a substantial influence on accuracy in real-world setting