Cure My FEVER: Building, Breaking, and Fixing Models for Fact-Checking

Christopher Hidey
Tuhin Chakrabarty
Tariq Alhindhi
Siddharth Varia
Kriste Krstovski
Mona Diab
Smaranda Muresan
Automated Fact-checking and Related Tasks

Source Trustworthiness  Fact-checking
### Automated Fact-Checking

#### Datasets and Problem Formulation

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Source</th>
<th>Size</th>
<th>Input</th>
<th>Output</th>
<th>Evidence Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truth of Varying Shades</td>
<td>Rashkin et al. (2017)</td>
<td>74K</td>
<td>Claim sentences</td>
<td>6 truth levels</td>
<td>No evidence</td>
</tr>
<tr>
<td></td>
<td>Politifact + news websites</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIAR</td>
<td>Wang (2017)</td>
<td>12.8K</td>
<td>Claim Sentences</td>
<td>6 truth levels</td>
<td>metadata</td>
</tr>
<tr>
<td>Emergent</td>
<td>Ferreira and Vlachos (2016)</td>
<td>300 claims 2,595 articles</td>
<td>Pair (claim, article headline)</td>
<td>for, against, observes</td>
<td>News Articles</td>
</tr>
<tr>
<td></td>
<td>Snopes.com Twitter</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FNC-1</td>
<td>Pomerleau and Rao (2017)</td>
<td>50K</td>
<td>Pair (headline, article body)</td>
<td>agree, disagree, discuss, unrelated</td>
<td>News Articles</td>
</tr>
<tr>
<td></td>
<td>Emergent</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Automated Fact-Checking Datasets and Problem Formulation

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Source</th>
<th>Size</th>
<th>Input</th>
<th>Output</th>
<th>Evidence Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truth of Varying Shades</td>
<td>Rashkin et al. (2017)</td>
<td>74K</td>
<td>Claim sentences</td>
<td>6 truth levels</td>
<td>No evidence</td>
</tr>
<tr>
<td></td>
<td>Politifact + news websites</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIAR</td>
<td>Wang (2017)</td>
<td>12.8K</td>
<td>Claim Sentences</td>
<td>6 truth levels</td>
<td>metadata</td>
</tr>
<tr>
<td></td>
<td>Politifact</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emergent</td>
<td>Ferreira and Vlachos (2016)</td>
<td>300 claims</td>
<td>Pair (claim, article headline)</td>
<td>for, against, observes</td>
<td>News Articles</td>
</tr>
<tr>
<td></td>
<td>Snopes.com Twitter</td>
<td>2,595 articles</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FNC-1</td>
<td>Pomerleau and Rao (2017)</td>
<td>50K</td>
<td>Pair (headline, article body)</td>
<td>agree, disagree, discuss, unrelated</td>
<td>News Articles</td>
</tr>
<tr>
<td></td>
<td>Emergent</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Synthetic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Overview

- FEVER: Fact Extraction and VERification of 185,445 claims
- Dataset
  - Claim Generation
  - Claim Labeling
- System
  - Document Retrieval
  - Sentence Selection
  - Textual Entailment

Claim: The Rodney King riots took place in the most populous county in the USA.

[wiki/Los_Angeles_Riots]
The 1992 Los Angeles riots, also known as the Rodney King riots were a series of riots, lootings, arsons, and civil disturbances that occurred in Los Angeles County, California in April and May 1992.

[wiki/Los_Angeles_County]
Los Angeles County, officially the County of Los Angeles, is the most populous county in the USA.

Verdict: Supported
Claim Generation

- Sample sentences from the introductory section of 50,000 popular pages (5,000 of Wikipedia’s most accessed pages and their linked pages)

- **Task:** given a sample sentence, generate a set of claims containing a single piece of information focusing on the entity that its original Wikipedia page was about.
  - Entities: a dictionary of terms with wikipedia pages.
  - Create mutations of the claims.
  - Average claim length is 9.4 tokens
Claim Labeling

- In 31.75% of the claims more than one sentence was considered appropriate evidence.
- Claims require composition of evidence from multiple sentences in 16.82% of cases.
- In 12.15% of the claims, this evidence was taken from multiple pages.
- IAA in evidence retrieval 95.42% precision and 72.36% recall.

Table 1: Dataset split sizes for SUPPORTED, REFUTED and NOTENOUGHINFO (NEI) classes
Given a factual claim involving one or more entities:

- Extract textual evidence (set of sentences) that could support or refute the claim.

Relevant Documents:

```
Shane Lee Lindstrom (born February 11, 1994), known by the stage name Murda Beatz, is a Canadian hip hop record producer and songwriter from Fort Erie, Ontario. He is noted for producing songs such as “No Shopping” by rapper French Montana and “Back on Road” by rapper Gucci Mane[1]; Murda has also produced several tracks for various artists such as...
```

Candidate Evidence Sentences:

```
Shane Lee Lindstrom (born February 11, 1994), known by the stage name Murda Beatz, is a Canadian hip hop record producer and songwriter from...
```

Prediction: REFUTED

"Murda Beatz’s real name is Marshall Mathers."
DATA AND METRICS

- 185,445 Claims

<table>
<thead>
<tr>
<th>Split</th>
<th>SUPPORTED</th>
<th>REFUTED</th>
<th>NEI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>80,035</td>
<td>29,775</td>
<td>35,639</td>
</tr>
<tr>
<td>Dev</td>
<td>6,666</td>
<td>6,666</td>
<td>6,666</td>
</tr>
<tr>
<td>Test</td>
<td>6,666</td>
<td>6,666</td>
<td>6,666</td>
</tr>
</tbody>
</table>

Metric:

- FEVER score = label accuracy conditioned on providing at least one complete set of evidence
Relevant Documents

- Google API: retrieve top documents for the claim
- Wikipedia API: Retrieve top documents for each named entity in the claim
- Query Wikipedia Search API with the subject of the claim

Candidate Evidence Sentences

- Use contextualized word embeddings (ELMO) to represent the claim and candidate evidence sentences.
- Compute cosine similarity and retrieve the top 5 most relevant sentences from the relevant documents

Textual Entailment Task

- Model each Claim – Candidate Evidence pair separately
- Do on top 3 candidates

Ranked 6th on the task last year on FEVER score
## RESULTS FOR ALL STAGES

### Doc retrieval

<table>
<thead>
<tr>
<th>METHOD</th>
<th>AVG K</th>
<th>COVERAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOOGLE API</td>
<td>2</td>
<td>79.5%</td>
</tr>
<tr>
<td>NER</td>
<td>2</td>
<td>77.1%</td>
</tr>
<tr>
<td>DEPENDENCY PARSE</td>
<td>1</td>
<td>80.0%</td>
</tr>
<tr>
<td>COMBINED</td>
<td>3</td>
<td>94.4%</td>
</tr>
<tr>
<td>THORNE ET AL (2017)</td>
<td>5</td>
<td>55.3%</td>
</tr>
</tbody>
</table>

### Evidence Recall

<table>
<thead>
<tr>
<th>DATASET</th>
<th>RECALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHARED TASK DEV</td>
<td>78.40</td>
</tr>
<tr>
<td>BLIND TEST SET</td>
<td>75.89</td>
</tr>
</tbody>
</table>

### Entailment Accuracy

<table>
<thead>
<tr>
<th>DATASET</th>
<th>ACCURACY</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHARED TASK DEV</td>
<td>58.77</td>
</tr>
<tr>
<td>BLIND TEST SET</td>
<td>57.45</td>
</tr>
</tbody>
</table>

### FEVER score

<table>
<thead>
<tr>
<th>DATA</th>
<th>PIPELINE</th>
<th>FEVER</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEV</td>
<td>THORNE ET AL (2018)</td>
<td>31.27</td>
</tr>
<tr>
<td></td>
<td>OURS</td>
<td>50.83</td>
</tr>
<tr>
<td>TEST</td>
<td>THORNE ET AL (2018)</td>
<td>27.45</td>
</tr>
<tr>
<td></td>
<td>OURS</td>
<td>49.06</td>
</tr>
</tbody>
</table>
ERROR ANALYSIS

- System wrongly penalized for not matching gold evidence

Claim: Aristotle spent time in Athens

System Prediction (correct): Supported

System Evidence (not in gold): At seventeen or eighteen years of age, he joined Plato’s Academy in Athens and remained there until the age of thirty-seven

System Evidence (not in gold): Shortly after Plato died, Aristotle left Athens and at the request of Philip II of Macedon, tutored Alexander the Great beginning in 343 BC
ERROR ANALYSIS

- Need better semantics (to distinguish NotEnoughInfo from Supported)

Claim: *Happiness in Slavery* is a *gospel* song by Nine Inch Nails

System Prediction: Supported

Gold Label: NotEnoughInfo

System Evidence: *Happiness in Slavery*, is a song by American industrial rock band Nine Inch Nails from their debut extended play (EP), Broken(1992)
Fact Extraction and VERification (FEVER) Version 2

Breakers

Development of adversarial claims

Builders & Fixers

Development of initial system and targeted improvements
1) Multiple propositions: Claims that require multi-hop document or sentence retrieval
   a) CONJUNCTION

   Janet Leigh was from New York. Janet Leigh was an author.

   -> Janet Leigh was from New York and was an author.
Breakers

1) Multiple propositions
   a) CONJUNCTION
   b) MULTI-HOP REASONING

The Nice Guys is a 2016 action comedy film. -> The Nice Guys is a 2016 action comedy film directed by a Danish screenwriter known for the 1987 action film Lethal Weapon.

[The_Nice_Guys]

[Shane_Black]
Breakers

1) Multiple propositions
   a) CONJUNCTION
   b) MULTI-HOP REASONING
   c) ADDITIONAL UNVERIFIABLE PROPOSITIONS

Duff McKagan is an American citizen

-> Duff McKagan is an American citizen born in Seattle.
Breakers

1) Multiple propositions
   a) CONJUNCTION
   b) MULTI-HOP REASONING
   c) ADDITIONAL UNVERIFIABLE PROPOSITIONS

2) Temporal reasoning
   a) DATE MANIPULATION

   in 2001 -> in the first decade of the 21st century

   in 2009 → 3 years before 2012
Breakers

1) Multiple propositions
   a) CONJUNCTION
   b) MULTI-HOP REASONING
   c) ADDITIONAL UNVERIFIABLE PROPOSITIONS

2) Temporal reasoning
   a) DATE MANIPULATION
   b) MULTI-HOP TEMPORAL REASONING

The first governor of the Indiana Territory lived long enough to see it become a state.

Admittance of Indiana Territory (1816) BEFORE William Henry Harrison (death 1841)
Breakers

1) Multiple propositions
   a) CONJUNCTION
   b) MULTI-HOP REASONING
   c) ADDITIONAL UNVERIFIABLE PROPOSITIONS

2) Temporal reasoning
   a) DATE MANIPULATION
   b) MULTI-HOP TEMPORAL REASONING

3) Ambiguity and lexical variation
   a) ENTITY DISAMBIGUATION

   Patrick Stewart -> Patrick Maxwell Stewart
Breakers

1) Multiple propositions
   a) CONJUNCTION
   b) MULTI-HOP REASONING
   c) ADDITIONAL UNVERIFIABLE PROPOSITIONS

2) Temporal reasoning
   a) DATE MANIPULATION
   b) MULTI-HOP TEMPORAL REASONING

3) Ambiguity and lexical variation
   a) ENTITY DISAMBIGUATION
   b) LEXICAL SUBSTITUTION

   filming -> shooting
Builders

1a Candidate Document Selection

1) Google
2) NER
3) POS
Builders

1a Candidate Document Selection

1) Google
2) NER
3) POS

2 Sentence Ranking

3 Relation Prediction

Joint Pointer Network
Fixers

1a Candidate Document Selection

1) Google
2) NER
3) POS
4) TF-IDF

1b Document Ranking

Pointer Network

2 Sentence Ranking

Joint Pointer Network

3 Relation Prediction

Overgenerate and re-rank to handle ambiguity
Fixers

1a Candidate Document Selection

1) Google
2) NER
3) POS
4) TF-IDF

1b Document Ranking

Pointer Network

1) Google
2) NER
3) POS
4) TF-IDF

2 Sentence Ranking

Joint Pointer Network

3 Relation Prediction

Sequence prediction to handle multiple propositions

Overgenerate and re-rank to handle ambiguity
Fixers

1a Candidate Document Selection

1) Google
2) NER
3) POS
4) TF-IDF

1b Document Ranking

Pointer Network

2 Sentence Ranking

Joint Pointer Network

3 Relation Prediction

Sequence prediction to handle multiple propositions

Overgenerate and re-rank to handle ambiguity

Post-processing to handle temporal relations
Pointer Network

Claim  Evidence

c  e₀

c  e₁

c  e₂

c  e₃

Candidate sentences
Pointer Network

Model fine-tuned on gold claim and evidence pairs

Candidate sentences
**Claim**: The Nice Guys is a 2016 film directed by a Danish screenwriter known for Lethal Weapon.

**Evidence**:  
- e\(^0\) = The Nice Guys is a 2016 American neo-noir crime black comedy film directed by Shane Black …  
- e\(^1\) = Shane Black… is an American filmmaker… written such films as Lethal Weapon…  
- e\(^2\) = He made his directorial debut with the film Kiss Kiss Bang Bang…  

**Model fine-tuned on gold claim and evidence pairs**

**LSTM decoder**

**Concatenate evidence to make label prediction, train using RL**
c=The Nice Guys is a 2016 film directed by a Danish screenwriter known for Lethal Weapon.

e^0=Nice Guys

e^1=Shane Black

Candidate sentences or documents

Model fine-tuned on gold claim and evidence pairs

LSTM decoder
Pointer Network

Model fine-tuned on gold claim and evidence pairs

Candidate documents or sentences

c=The Nice Guys is a 2016 film directed by a Danish screenwriter known for Lethal Weapon.

e^0=The Nice Guys is a 2016 American neo-noir crime black comedy film directed by Shane Black ...

I^0=NEI

e^1=Shane Black... is an American filmmaker... written such films as Lethal Weapon...

I^1=REFUTES

e^2=He made his directorial debut with the film Kiss Kiss Bang Bang...

I^2=REFUTES
Post-processing for Temporal Relations

1. Extract temporal expressions:
   The Latvian Soviet Socialist Republic was a republic of the Soviet Union 3 years after 2009.
   1. Open IE -> 3 years after 2009
   2. Normalize -> 2012

2. Compare only dates in retrieved evidence:
   The Soviet Union … existed from 1922 to 1991.
   1991 < 2012 -> Refutes
# Results - Breakers

<table>
<thead>
<tr>
<th>Team</th>
<th>#</th>
<th>Raw Potency</th>
<th>Correctness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>498</td>
<td>60.34</td>
<td>82.33</td>
</tr>
<tr>
<td>NbAuzDrLqg</td>
<td>102</td>
<td>79.66</td>
<td>64.71</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td>501</td>
<td>68.51</td>
<td>81.44</td>
</tr>
<tr>
<td>TMLab</td>
<td>79</td>
<td>79.97</td>
<td>84.81</td>
</tr>
</tbody>
</table>
# Results - Builders

<table>
<thead>
<tr>
<th>Team</th>
<th>FEVER 1.0</th>
<th>FEVER 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Athene</td>
<td>61.58</td>
<td>25.35</td>
</tr>
<tr>
<td>UNC</td>
<td>64.21</td>
<td>30.47</td>
</tr>
<tr>
<td><strong>Builders</strong></td>
<td><strong>67.08</strong></td>
<td><strong>32.92</strong></td>
</tr>
<tr>
<td>Dominiks</td>
<td>68.46</td>
<td>35.82</td>
</tr>
<tr>
<td>UCL MR</td>
<td>62.52</td>
<td>35.83</td>
</tr>
<tr>
<td>Papelo</td>
<td>57.36</td>
<td>37.31</td>
</tr>
</tbody>
</table>
## Results - Fixers

<table>
<thead>
<tr>
<th>Team</th>
<th>FEVER 1.0</th>
<th>FEVER 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Athene</td>
<td>61.58</td>
<td>25.35</td>
</tr>
<tr>
<td>UNC</td>
<td>64.21</td>
<td>30.47</td>
</tr>
<tr>
<td><strong>Builders</strong></td>
<td>67.08</td>
<td>32.92</td>
</tr>
<tr>
<td>Dominiks</td>
<td><strong>68.46</strong></td>
<td>35.82</td>
</tr>
<tr>
<td>UCL MR</td>
<td>62.52</td>
<td>35.83</td>
</tr>
<tr>
<td><strong>Fixers</strong></td>
<td><strong>68.8</strong></td>
<td><strong>36.61</strong></td>
</tr>
<tr>
<td>Papelo</td>
<td>57.36</td>
<td><strong>37.31</strong></td>
</tr>
</tbody>
</table>
Questions?