DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking

Christopher Hidey, Tuhin Chakrabarty, Tariq Alhindi, Siddharth Varia, Kriste Krstovski, Mona Diab, Smaranda Muresan
Motivation and Background

- Why do we need fact-checking?
  - Misinformation/disinformation on the rise (Ireton, 2018)
  
  Journalists must examine “check-worthy” claims for truth (Hassan et al., 2017)

- Related Work
  - Fact-checking websites (Vlachos and Reidel, 2014; Alhindí et al., 2018)
  
  News articles (Pomerleau and Rao, 2017)
  
  Community forums (Mihaylova et al., 2018)
Challenges

1) Professional fact-checking organizations need to synthesize evidence from multiple sources (Graves, 2018)

2) Numerical comparatives alone are indicative of truthful statements (Rashkin et al., 2017) but temporal reasoning is challenging (Mirza and Tonnelli, 2016)

3) Fact-checking is sensitive to ambiguous entities (Thorne and Vlachos, 2018) or lexical features (Nakashole and Mitchell, 2014)
Towards Addressing these Challenges

Data

Development of adversarial claims

Methods

Development of fact-checking system with targeted improvements
Data

Murda Beatz’s real name is Marshall Mathers.

[Murda Beatz]

Shane Lee Lindstrom, known professionally as Murda Beatz ...

REFUTES
Data

FEVER 1.0 (Thorne et al. 2018)

FEVER 2.0 (Thorne et al. 2019)

- adversarial “attacks” submitted by participants
- evaluated against FEVER systems
Data: Adversarial Claims

1) Multiple propositions
   a) CONJUNCTION

   Janet Leigh was from New York. Janet Leigh was an author.
Data: Adversarial Claims

1) Multiple propositions
   a) CONJUNCTION

   Janet Leigh was from New York and Janet Leigh was an author.
Data: Adversarial Claims

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

[The Nice Guys]

The Nice Guys is a 2016 action comedy film.
Data: Adversarial Claims

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

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

[Shane Black]
Data: Adversarial Claims

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

   Duff McKagan is an American citizen.
Data: Adversarial Claims

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

   Duff McKagan is an American citizen born in Seattle.
Data: Adversarial Claims

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

2) Temporal reasoning
   a) DATE MANIPULATION

   Artpop was Gaga’s second consecutive number one record in the United States in 2009.
Data: Adversarial Claims

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

2) Temporal reasoning
   a) DATE MANIPULATION

   Artpop was Gaga’s second consecutive number one record in the United States in 2009 before 2010.
Data: Adversarial Claims

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

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

   [William Henry Harrison] ⛵️ [Indiana Territory] ⛵️

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

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

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

   [William Henry Harrison] 🌍  [Indiana Territory] 🌍

   The first governor of the Indiana Territory lived long enough to see it
   become a state.  
   (death 1841)
   (admittance 1816)
Data: Adversarial Claims

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

   Kate Hudson is a left wing political activist.

   [Kate Hudson] 🌍
Data: Adversarial Claims

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

Kate Hudson is a left wing political activist.

[Kate Hudson] [Kate Hudson (activist)]
Data: Adversarial Claims

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

   The Last Song began filming in 2009.
Data: Adversarial Claims

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

   The Last Song began filming **shooting** in 2009.

DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL. 2020.
Evaluation: Data

- Organizer Baseline
- Kim and Allen (2019)
- Ours
- Niewinski et al. (2019)
Evaluation: Data
Evaluation: Data

![Bar chart showing different models compared by inverse FEVER score.]

- Organizer Baseline
- Kim and Allen (2019)
- Niewinski et al. (2019)
- Multi-hop
- Multi-hop temporal
Evaluation: Data

![Bar chart showing comparison of different methods]

- Organizer Baseline
- Kim and Allen (2019)
- Ours
- Niewinski et al. (2019)
- Multi-hop
- Multi-hop temporal

**Inverse FEVER score**
Methods: Fact-Checking System

1a Candidate Document Selection

1) Google
2) TF-IDF
3) NER
4) POS
Methods: Fact-Checking System

1a Candidate Document Selection
   1) Google
   2) TF-IDF
   3) NER
   4) POS

1b Document Ranking
   Pointer Network

Overgenerate and re-rank to handle ambiguity

DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking. 
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL 2020.
Methods: Fact-Checking System

Candidate Document Selection

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

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

**Methods: Fact-Checking System**

1a Candidate Document Selection
- 1) Google
- 2) TF-IDF
- 3) NER
- 4) POS

1b Document Ranking
- Pointer Network

2 Sentence Ranking
- Joint Pointer Network

3 Relation Prediction

Overgenerate and re-rank to handle ambiguity

Post-processing to handle temporal relations

Sequence prediction to handle multiple propositions
Methods: Pointer Network

c: Michelle Obama’s husband was born in Kenya.
Methods: Pointer Network

Fine-tuned on gold claim and evidence pairs

Methods: Pointer Network

DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking. 
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL. 2020.
Methods: Pointer Network

DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
Hidey, Chakrabarty, Alhindi, Varia, Krstovski, Diab, Muresan. ACL. 2020.
Claim: Michelle Obama's husband was born in Kenya.

Evidence: Obama... born in Hawaii

NOT ENOUGH INFO

DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking.
Hidey, Chakrabarty, Alhind, Varia, Krstovski, Diab, Muresan. ACL. 2020.
Methods: Pointer Network

Joint Sentence Selection and Relation Prediction

Claim: Michelle Obama's husband was born in Kenya.

Evidence:
- e₀: Michelle...married Barack
  REFUTES
- e₁: Obama...born in Hawaii
  NOT ENOUGH INFO

Methods: Pointer Network

Claim: Michelle Obama’s husband was born in Kenya.

Evidence:
- $e_0$: Michelle…married Barack (REFUTES)
- $e_1$: Obama…born in Hawaii (NOT ENOUGH INFO)

Post-processing for temporal reasoning

Methods: Temporal Post-processing

1. Extract with Open IE and normalize:
   The Latvian Soviet Socialist Republic was a republic of the Soviet Union 3 years after 2009.

2. Compare only dates in retrieved evidence:
   The Soviet Union … existed from 1922 to 1991.
Evaluation: Data

![Bar chart showing inverse FEVER scores for different methods]

- Organizer Baseline
- Kim and Allen (2019)
- Ours
- Niewinski et al. (2019)
- Multi-hop
- Multi-hop temporal
Evaluation: Methods

Performance by System

- Malon (2018)
- Baseline
  - no doc re-ranking
  - no relation sequence
  - no temporal post-proc
- Stammbach and Neumann (2019)
- Our System
Evaluation: Methods

Performance by System

Evidence Recall

FEVER 1.0

Nie et al. (2019)
Yoneda et al. (2018)
Hanselowski et al. (2018)
Our System
Evaluation: Attack Types

Performance by Attack Type

- Baseline: no doc re-ranking, no relation sequence, no temporal post-proc
- Our System
Evaluation: Attack Types

Performance by Attack Type

- Baseline
  - no doc re-ranking
  - no relation sequence
  - no temporal post-proc

- Our System

Evidence Retrieval

<table>
<thead>
<tr>
<th>Attack Type</th>
<th>Baseline</th>
<th>Our System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-hop</td>
<td>30</td>
<td>50</td>
</tr>
<tr>
<td>Multi-hop Temp.</td>
<td>10</td>
<td>20</td>
</tr>
</tbody>
</table>
Evaluation: Qualitative Analysis

Baseline:

Honeymoon is a major-label record by Elizabeth Woolridge Grant.
Evaluation: Qualitative Analysis

Our System:

Honeymoon is a major-label record by Elizabeth Woolridge Grant.

[Honeymoon (Lana del Rey album)] [Lana del Rey]

SUPPORTS
Evaluation: Qualitative Analysis

Baseline:

The MVP of the 1976 Canada Cup tournament was born before the tournament was first held.

[✓] [Bobby Orr] [World Cup of Hockey] [Canada Cup]

[✗] [REFUTES]
Evaluation: Qualitative Analysis

Our System:

The MVP of the 1976 Canada Cup tournament was born before the tournament was first held.

[✓] [✓] [Bobby Orr] [1976 Canada Cup]

[✗] NOT ENOUGH INFO
Questions?

Come join our discussion slot(s):

Wednesday July 8, 2020 14B NLP Applications-11 18:00 UTC+0

Wednesday July 8, 2020 15B NLP Applications-12 21:00 UTC+0
# Evaluation: Ablation

<table>
<thead>
<tr>
<th>System</th>
<th>FEVER 1.0 Dev</th>
<th>FEVER 2.0 Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Builders</strong></td>
<td>69.2</td>
<td>36.2</td>
</tr>
<tr>
<td>Ptr+rel.seq.loss</td>
<td>73.17</td>
<td>39.86</td>
</tr>
<tr>
<td>+doc. rank</td>
<td>-</td>
<td>41.91</td>
</tr>
<tr>
<td>+dateProc <strong>(Fixers)</strong></td>
<td>-</td>
<td>43.36</td>
</tr>
</tbody>
</table>
## Evaluation: Data

<table>
<thead>
<tr>
<th>Team</th>
<th>#</th>
<th>Potency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>498</td>
<td>49.68</td>
</tr>
<tr>
<td>Kim and Allen (2019)</td>
<td>102</td>
<td>51.54</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td>501</td>
<td>55.79</td>
</tr>
<tr>
<td>Niewinski et al. (2019)</td>
<td>79</td>
<td>66.83</td>
</tr>
<tr>
<td><strong>Multi-hop</strong></td>
<td>188</td>
<td><strong>90.34</strong></td>
</tr>
<tr>
<td><strong>Multi-hop Temporal</strong></td>
<td>55</td>
<td><strong>98</strong></td>
</tr>
</tbody>
</table>
## Evaluation: Methods

<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>Our Baseline - RL</strong></td>
<td><strong>67.08</strong></td>
<td><strong>32.92</strong></td>
</tr>
<tr>
<td>Dominiks</td>
<td><strong>68.46</strong></td>
<td><strong>35.82</strong></td>
</tr>
<tr>
<td>UCL MR</td>
<td>62.52</td>
<td>35.83</td>
</tr>
<tr>
<td><strong>Our System</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>